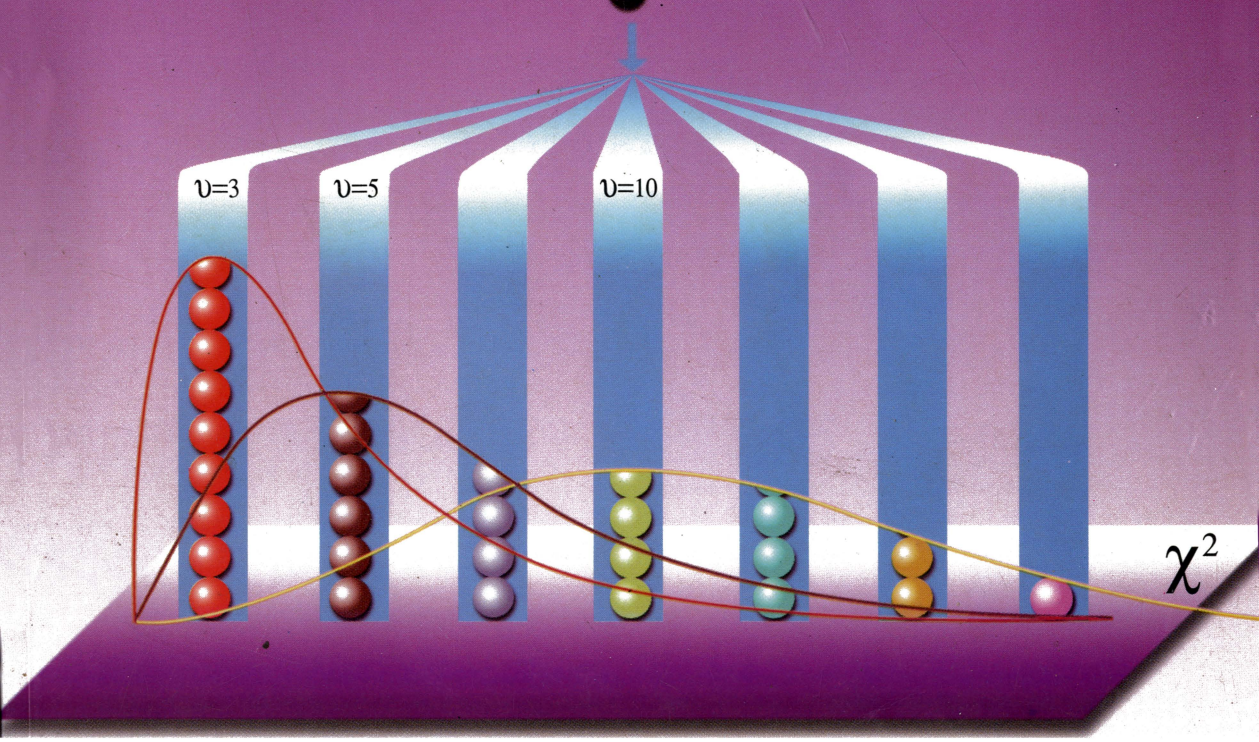
Rejeb HAJI

Statistique Appliquée



Statistique

Appliquée

A Faouzia, mon épouse
&
à Amel & Hager mes deux filles

DU MEME AUTEUR :

STATISTIQUE & PROBABILITES
(*Manuel publié en Janvier 2000*)

• *Toute reproduction faite sans le consentement de l'auteur ou de ses ayants droits est illicite et constitue une contrefaçon réprimée par la loi tunisienne.*

Rejeb HAJI

Statistique Appliquée

- Rappels théoriques
- Auto-évaluation
- Etudes de cas

• L'œuvre scientifique vaut par l'intérêt qu'elle suscite. C'est pourquoi, je me fais un devoir, ici, de remercier la *Fondation Universitaire BOUEBDELLI* pour tout l'intérêt qu'elle porte à la production scientifique tunisienne et à son encouragement.

INTRODUCTION

Dans le manuel « ***Statistique et Probabilités*** » nous avons montré comment symboliser les observations recueillies sur une population par des représentations graphiques puis comment les résumer au moyen de paramètres de position, de paramètres de dispersion et de paramètres de forme. Nous avons, également, introduit les concepts du calcul des probabilités et traité de quelques modèles empiriques qui permettent de mieux comprendre les phénomènes et de les analyser.

Ce deuxième volume complète les lois décrites dans le premier manuel par d'autres lois théoriques d'usage courant, considérées comme outils techniques en statistiques. Nous présentons, également, dans une partie théorique la notion de convergence de variables et les théorèmes y afférents pour une analyse concrète de la réalité. Enfin, nous abordons un ensemble de méthodes qui permettent de juger des propriétés des paramètres relatifs à une population insuffisamment connue, au moyen d'une partie prélevée au hasard au sein de cette population. Ces méthodes sont souvent désignées sous l'appellation ***d'inférence statistique*** ou de ***statistique inductive*** ou encore de ***statistique mathématique.***

Pour l'élaboration de ce deuxième manuel, comme pour le premier, nous avons consulté, bien entendu, la littérature classique citée en référence mais, nous nous sommes référés aussi à notre propre expérience aussi bien dans les domaines de l'activité économique que dans l'enseignement. Pour faciliter la compréhension des concepts développés, nous avons choisi de conclure les chapitres par une sélection de questions d'***auto-évaluation*** et d'***études de cas*** tirés de champs d'application différents.

Ces exercices, diversifiés, ont été empruntés à des séries d'examens ou fournis par Mme Haji, ancienne assistante hospitalo-universitaire et actuellement médecin des hôpitaux de Tunis.

L'objectif poursuivi est double : faire mieux comprendre les notions de base évoquées dans le chapitre afin de contrôler les connaissances acquises et permettre du même coup, à un auditoire le plus large possible de se familiariser avec une science qui se développe de jour en jour dans tous les domaines de la vie.

La méthodologie statistique, faut-il le souligner, n'a pas de frontières ni de particularités. Elle s'applique largement à l'économie et à toutes les sciences sociales et de la nature.

Qu'il me soit permis de réitérer mes remerciements à mon épouse Faouzia pour m'avoir fourni des applications relevant du domaine médical et guidé par ses conseils judicieux dans la mise en page et la présentation pédagogique de cet ouvrage. Mes remerciements s'adressent également à tous ceux qui m'ont aidé directement ou indirectement à mener à bien ce travail.

Toute œuvre, quelque soin qu'on y apporte, demeure entachée d'imperfections. Aussi serais-je reconnaissant à tous ceux qui par leurs remarques et critiques ou par les erreurs qu'ils m'auraient signalées, me permettront de l'améliorer. J'y verrais une marque d'intérêt pour un domaine appelé à occuper de plus en plus de place dans nos économies et sociétés et dont la maîtrise est synonyme de rationalité, d'anticipation et de progrès.

R.HAJI
Septembre 2001

TABLE DES MATIERES

Chapitre 1 : Variable aléatoire • Lois de probabilité

I- Variable aléatoire 2
- 1- Définitions 2
- 2- Valeurs typiques 2

II- Convergences • Inégalités 6
- 1- Convergences 6
- 2- Inégalités 6

III- Théorèmes fondamentaux 8
- 1- Loi des grands nombres 9
- 2- Théorème central limite 9

IV- Fonctions génératrices 10
- 1- Définitions 10
- 2- Propriétés 10
- 3- Applications 11

V- Lois dérivées 12
- 1- Généralisation de la loi uniforme 13
- 2- Loi gamma 14
- 3- Loi log-normale 16
- 4- Loi du χ^2 16
- 5- Loi de Student 19
- 6- Loi de Fisher-Snedecor 20

Auto-évaluation 24
Exercices 27
Corrigés 29

Chapitre 2 : Echantillonnage

I- Echantillonnage 38
- 1- Introduction 38
- 2- Sondage • Recensement 39

II- Construction d'un échantillon 42
- 1- Méthodes empiriques 42

2- Méthodes aléatoires 45
III- Distribution d'échantillonnage 51
1- Exemple 51
2- Généralisation 52
IV- Distribution des paramètres 55
1- Théorèmes 55
2- Distribution d'une fréquence 55
3- Distribution de la moyenne 56
4- Distribution de la variance 57
V- Intervalle de Pari 58
1- Définition 58
2- Construction 58

Auto-évaluation 59
Exercices 60
Corrigés 63

Chapitre 3 : Estimation ponctuelle
I- Terminologie • Définitions 70
II- Classification des estimateurs 71
1- Estimateur convergent 71
2- Estimateur sans biais 73
3- Estimateur efficace 74
4- Estimateur exhaustif 76
III- Méthodes d'estimation 78
1- Méthode des moments 78
2- Méthode des moindres carrés 80
3- Méthode du maximum de vraisemblance 80
IV- Estimation des paramètres 85
1- Estimation d'une proportion 85
2- Estimation d'une moyenne 86
3- Estimation de la variance 87

Auto-évaluation 89
Exercices 91
Corrigés 93

Chap-itre 4 : Intervalle de confiance

I- Intervalle pour une proportion 102
1- Intervalle pour p 102
2- Intervalle pour une différence de proportions 106
II- Intervalle pour une moyenne 107
1- Distribution quelconque 107
2- Distribution normale 109
Intervalle pour une différence de moyennes 110
1- Distribution quelconque 110
2- Distribution normale 111
III- Intervalle pour une variance 113
1- Distribution quelconque 113
2- Distribution normale 114
3- Distribution approximativement normale 115
Intervalle pour une différence de variances 116
1- Distribution quelconque 117
2- Distribution approximativement normale 117
IV- Coefficient de corrélation 118

Auto-évaluation 120
Exercices 121
Corrigés 124

Chapitre 5 : Tests paramétriques

I- Principes généraux 136
1- Terminologie 136
2- Méthodes 137
3- Applications 140
II- Tests de proportions 142
1- Test sur une proportion 142
2- Test de comparaisons 146
III- Tests de moyennes 149
1- Test sur une moyenne 149
2- Test de comparaisons 151

IV- Tests de variances 153
1- Test sur une variance 153
2- Test de comparaisons 155
V- Test sur le coefficient de corrélation 164

Auto-évaluation 167
Exercices 169
Corrigés 174

Chapitre 6 : Tests non-paramétriques

I- Test du KHI-DEUX 190
1- Test d'ajustement 190
2- Test d'homogénéité 195
3- Test d'indépendance 196
II- Test des signes 197
1- Test sur une moyenne 197
2- Test de comparaisons 198
III- Test des rangs 199
1- Echantillons indépendants 199
2- Echantillons appariés 200
IV- Test des séquences 200

Auto-évaluation 202
Exercices 204
Corrigés 207

ANNEXES 215
SYNTHESE DES LOIS USUELLES 216
BIBLIOGRAPHIE SOMMAIRE 223
DES NOMS & DES DATES 224
INDEX ALPHABETIQUE 227
TABLES STATISTIQUES 235

Chapitre 1

VARIABLE ALEATOIRE
◆
LOIS DE PROBABILITE

☛ *Objectifs*

⇨ rappeler les inégalités de Markov et de Bienaymé-Tchebychev

⇨ étudier les modes de convergence

⇨ présenter le théorème central-limite

⇨ introduire la fonction génératrice

☛ *Mots clés*

Variable aléatoire – inégalité de Markov – inégalité de Bienaymé-Tchebychev – convergence en loi – convergence en moyenne quadratique – loi des grands nombres – théorème central-limite – fonction génératrice – fonctions génératrices des lois usuelles – lois : uniforme, gamma, log-normale, Khi-deux, Student, Fisher-Snedecor.

I *Variable aléatoire ◆ Loi de probabilité*

II *Inégalités et convergences*

III *Théorèmes fondamentaux*

IV *Fonctions génératrices*

V *Lois dérivées*

AUTO-EVALUATION & EXERCICES CORRIGES

Il s 'agit de rappeler dans ce qui suit la notion de variable aléatoire et de décrire les lois de probabilités dérivées des lois continues usuelles.

I - VARIABLE ALEATOIRE

Le rappel de certaines propriétés des variables aléatoires s'impose. Pour plus de détails, se référer au manuel « ***Statistique & Probabilités*** ».

I - 1 DEFINITIONS

Les *variables* sont les *caractères d'une population que l'on mesure*. Elles peuvent être :

- ♦ qualitatives (variables nominales)
- ♦ quantitatives (variables métriques et ordinales)
  - ♣ discrètes (comptées)
  - ♣ continues (mesurées).

Plus généralement :

Etant donné un univers fini Ω, toute application de Ω dans R est appelée **variable aléatoire** (notée ***v. a.)*** ou **aléa numérique.**

I - 2 VALEURS TYPIQUES

Les valeurs typiques peuvent se classer en :

- *paramètres de tendance centrale* : mode, médiane, moyenne...
- *paramètres de dispersion* : écart quadratique, écart-type, moments...
- *paramètres de forme* : coefficient de Pearson, coefficient de Yule...

I - 2. 1 Paramètres de tendance centrale

Ce sont les valeurs de la variable situées au «centre » de la distribution.

① Mode (dominante)

➢ ***Pour une loi discontinue*** :

On appelle *mode* la valeur de la variable qui correspond à la *plus grande probabilité*, c'est à dire la *valeur la plus probable*.

➢ ***Pour une loi continue*** :

On appelle *mode* (s) [*dominante*(s)] la ou les valeurs de la variable x où la fonction *f(x) densité de probabilité* (***d. d. p.***) présente un *maximum*.

Remarques

☞ Si le mode est unique, on dit que la loi est *unimodale*

☞ Sinon, elle est *plurimodale*.

② **Médiane (valeur équiprobable)**

On appelle *médiane* (ou *valeur équiprobable*) d'une distribution l'abscisse du point qui partage en *deux parties égales* l'effectif total.

③ **Espérance mathématique (moyenne)**

La *moyenne* ou *l'espérance mathématique* d'une variable aléatoire est notée $E(X)$.

➢Pour une **v. a. discrète** : $E(X) = \sum_{i \in N} p_i x_i$

➢Pour une **v. a. continue** : $E(X) = \int_{-\infty}^{+\infty} xf(x)dx$

Remarque

☞ L'espérance mathématique *existe* si la série est *absolument convergente.*

❑ ***Propriétés de l'espérance***

❶ $E(aX+b) = aE(X)+b$

❷ *Règle d'addition* :

$$E(\sum_{i \in N} X_i) = \sum_{i \in N} E(X_i)$$

❸ *Règle de multiplication*

Si les variables aléatoires sont *indépendantes*, on a

$E(X_1 \times X_2 \times \ldots \times X_n) = E(X_1) \times E(X_2) \times \ldots \times E(X_n)$

❹ Dans le cas d'une fonction à **plusieurs variables** aléatoires

$Z = \varphi(X, Y)$, l'espérance est définie par :

➢Pour **une loi discrète** :

$$E(Z) = E[\varphi(X,Y)] = \sum_{i,j} p_{ij}\varphi(x_i, y_j)$$

➢Pour **une loi continue** :

$$E(Z) = E[\varphi(X,Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \varphi(x,y) f(x,y)dxdy$$

I - 2. 2 Paramètres de dispersion

①Ecart absolu moyen (EAM)

On appelle *écart absolu moyen l'espérance mathématique* des *valeurs absolues des écarts* de la variable aléatoire X par rapport à une valeur a initialement choisie :

➢Pour **une v. a. discrète** : $E|X-a| = \sum_{i=1}^{n} p_i |x_i - a|$

➢Pour **une v. a. continue** : $E|X-a| = \int_{-\infty}^{+\infty} |x-a| f(x)dx$

②Ecart-type (écart quadratique moyen)

♦On appelle **variance** ou **fluctuation** (en anglais *standard deviation*) l'expression :

➢ Pour **une v. a. discrète**

$$\sigma^2 = Var(X) = E[X - E(X)]^2 = \sum_{1}^{n} [x_i - E(X)]^2 p_i$$

➢ Pour **une v. a. continue**

$$\sigma^2 = Var(X) = \int_a^b [x - E(X)]^2 f(x)dx \quad \text{avec } x \in [a; b]$$

❑ ***Propriétés de la variance***

❶ $Var(X) = E(X^2) - [E(X)]^2$

❷ $Var(aX+b) = a^2 Var(X)$

❸ Si on pose $a = \frac{1}{\sigma}$ et $b = -\frac{\mu}{\sigma}$ alors $E(\frac{x-\mu}{\sigma}) = 0$ et $V(\frac{x-\mu}{\sigma}) = 1$

la variable $T = \frac{x-\mu}{\sigma}$ est appelée **variable centrée réduite**.

❹ *Règle d'addition* : Si les variables aléatoires sont *indépendantes* alors

$$V(\sum_{i \in N} X_i) = \sum_{i \in N} V(X_i)$$

❺ Pour une combinaison linéaire à coefficients constants de *n v.a. indépendantes* :

$$V(\sum_{i \in N} a_i X_i) = \sum_{i \in N} a^2_i V(X_i)$$

☞ ***Cas particulier***

Si les ***n v. a.*** *indépendantes* ont *même moyenne* μ et *même écart- type* σ et si on considère leur moyenne arithmétique, on a :

$$\overline{X} = \frac{\sum_{1}^{n} X_i}{n} \quad \text{alors} \quad E(\overline{X}) = \mu \text{ et } V(\overline{X}) = \frac{\sigma^2}{n}$$

♦ On appelle **écart-type** ou **écart moyen quadratique centré** la racine carrée positive de la variance

$$\sigma^2=\sqrt{E(X^2)-\mu^2} \quad \text{avec} \quad E(X) = \mu$$

③ Les moments

Soit *r un entier positif ou nul*, on appelle :

✓ **moments non centrés** les moments d'ordre r rapportés à l'origine

$$\boxed{m_r = E(X^r)}$$

✓ **moments centrés** les moments d'ordre r rapportés à la moyenne

$$\boxed{\mu_r = E[(X-\mu)^r]}$$

☞ ***Cas particuliers*** :

$$\mu_0 = 1 \; ; \qquad \mu_1 = E(X-\mu) = 0$$

$$\mu_2 = \sigma^2 = m_2 - m_1^2$$

On obtient aisément les relations existant entre les moments centrés et les moments non centrés, soit :

$$\mu_3 = m_3 - 3\, m_1 m_2 + 2\, m^3$$

$$\mu_4 = m_4 - 4\, m_1 m_3 + 6 m_1^2 m_2 - 3 m^4$$

④ Autres valeurs typiques de dispersion

On peut définir d'autres paramètres pour caractériser la dispersion :

✓ **l'étendue** : la *différence* entre la *plus grande* et la *plus petite* des *valeurs observées* : $max\ x_i$ - $min\ x_i$ *pour* $1 \le i \le n$

✓ **l'amplitude** : la différence $b - a$ lorsque $X \in [a\ ;\ b]$

✓ **les quantiles** d'ordre α $(1 \le \alpha \le n)$ *notés* x_α.

En désignant par F la *fonction de répartition* de la variable aléatoire X, on a :

$$\boxed{F(x_\alpha) = \alpha \Leftrightarrow Prob(X \le x_\alpha) = \alpha}$$

On utilise couramment les *quartiles* Q_1, Q_2, Q_3 qui sont les quantiles d'ordre respectif : $\frac{1}{4}; \frac{1}{2}; \frac{3}{4}$ Ces quartiles partagent les observations en quatre parties égales.

✓ **l'intervalle interquartile** :

$$\boxed{Q_3 - Q_1}$$

Remarque

☞ Pour des raisons de commodité, on utilise *l'écart-type* ou la *variance* comme *valeur typique de dispersion*.

I - 2. 3 Paramètres de forme

Ils permettent de *repérer la forme* de la distribution. On choisit comme mesure du degré d'asymétrie l'un des moments, μ_3 par exemple, et on définit :

- *Coefficient d'asymétrie de Fisher* :

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \qquad \text{avec } \sigma \neq 0$$

$\gamma_1 = 0 \Rightarrow$ distribution *symétrique*

$\gamma_1 < 0 \Rightarrow$ distribution unimodale *étalée vers la gauche*

$\gamma_1 > 0 \Rightarrow$ distribution unimodale *étalée vers la droite*

- *Coefficient d'aplatissement de Fisher* :

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 \qquad \text{avec } \sigma \neq 0$$

$\gamma_2 = 0$ pour une *distribution normale*

Remarques

☞ Ces *coefficients* sont *invariants* par *changement d'origine* et *d'échelle*.

☞ Ils sont *sensibles* aux variations *d'échantillonnage*.

II - CONVERGENCES ◆ INÉGALITÉS.

Dans la pratique, on fait correspondre des épreuves aléatoires à des variables aléatoires. La notion de convergence de ces variables, très utilisée par la suite, *lie la statistique aux probabilités*. C'est par elle que se trouve justifiée l'utilisation du *Calcul des Probabilités* à l'analyse des phénomènes réels.

II - 1 CONVERGENCES

Nous allons rappeler, dans ce qui suit, un ensemble de *définitions* et de *théorèmes*, souvent utilisés dans l'étude du comportement asymptotique d'une suite infinie de variables aléatoires.

II - 1. 1 Convergence en probabilité

① **Définitions**

❶On dit qu'une suite de *variables aléatoires* $\{y_n\}$ tend en probabilité vers une *valeur certaine a* si :

$$\forall \varepsilon > 0 \quad \forall \eta > 0 \quad \exists N(\varepsilon, \eta): n > N \Rightarrow \text{Prob}\{ |y_n - a| > \varepsilon \} < \eta$$

Cette définition est une généralisation de la convergence habituelle en mathématiques.

❷Une suite de variables aléatoires $\{X_n\}$ tend vers une variable aléatoire X si la suite de v. a. $\{X_n - X\}$ tend en probabilité vers zéro.

$$\text{Prob}(|X_n - X| < \varepsilon) = 0 \quad \forall \varepsilon > 0$$

Cela signifie que X_n *se rapproche de X quand n augmente.*

② **Théorèmes**

❶*Soit* $\{X_n\}$ *une suite de variables aléatoires qui converge en probabilité vers une variable aléatoire X. Soit g une fonction définie et continue sur l'ensemble des réels R, on a :* $g(X_n) \to g(X)$

En particulier si a est une valeur certaine : $X_n \to a \Rightarrow g(X_n) \to g(a)$

❷*Soit* $\{X_n\}$ *une suite de variables aléatoires telle que :*

$E(X) \to \mu$ et $V(X) \to 0$ quand n $\to \infty$, *alors la suite converge en probabilité vers* μ.

II - 1. 2 Convergence en loi

① **Définition**

$\{X_n\}$ une suite de variables aléatoires *converge en loi* vers une variable aléatoire X, de fonction de répartition F, *lorsque* $n \to \infty$ si en tout point de continuité x de F, les fonctions de répartition F_n *tendent vers* F

$$\lim F_n(x) \to F(x) \qquad \text{quand } n \to \infty$$

② **Théorème**

La *convergence en probabilité implique* la *convergence en loi.*

Remarque

☞ La réciproque n'est vraie que si la convergence se fait vers une variable certaine.

II - 1. 3 Convergence en moyenne quadratique

On suppose que les *variables aléatoires admettent un moment de second ordre fini.*

① **Définitions**

❶ On définit sur l'ensemble des variables aléatoires, ayant une espérance et une variance finies, *une norme* par l'expression :

$$\|X\| = \sqrt{E(|X|^2)}$$

❷ $\{X_n\}$ une suite de variables aléatoires *converge en moyenne quadratique* vers une variable aléatoire X si $\|X_n - X\| \to 0 \quad (n \to \infty)$

② **Théorème**

La convergence en moyenne quadratique entraîne la convergence en probabilité.

Remarque

☞ La réciproque n'est pas vraie.

II - 2 INÉGALITÉS

Deux inégalités, sont souvent employées, soit pour majorer des probabilités, soit pour démontrer des théorèmes :

II - 2. 1 Inégalité de Markov

Soit X une variable aléatoire définie sur un espace probabilisé (Ω, p) tel que $X(\Omega) \subset R_+^*$ alors pour tout $a > 0$ $\quad \text{Prob}(X \geq a) \leq \dfrac{E(X)}{a}$

où $E(X)$ est l'espérance de X avec $E(X) = \int_0^{+\infty} f(x)dx$ puisque $X(\Omega) \subset R_+^*$

Comme $a > 0 \;\; a \leq x \Rightarrow af(x) \leq xf(x)$

$\Rightarrow \int_a^{+\infty} a\, f(x)dx \leq \int_0^{+\infty} x\, f(x)dx$ d'où

$$\boxed{a\,P(X \geq a) \leq E(X) \Rightarrow P(X \geq a) \leq \frac{E(X)}{a}}$$

II - 2. 2 Inégalité de Bienaymé-Tchebychev

Soit X une variable aléatoire qui a pour espérance $E(X)$ et pour variance $V(X)$ différente de zéro, l'inégalité de Bienaymé-Tchebychev est donnée par :

$$\forall \varepsilon > 0 \quad P(|X - E(X)| > \varepsilon) \leq \frac{V(X)}{\varepsilon^2}$$

$\forall \varepsilon > 0 \quad \text{Prob}(|X - E(X)| > \varepsilon) \Rightarrow \text{Prob}(Y^2 > \varepsilon^2)$

d'après l'inégalité de Markov

$$\text{Prob}(Y^2 > \varepsilon^2) \leq \frac{E|X - E(X)|^2}{\varepsilon^2} = \frac{V(X)}{\varepsilon^2}$$

III - THEOREMES FONDAMENTAUX

Il s'agit d'énoncer *deux théorèmes fondamentaux* qui vont nous permettre de *contourner la difficulté* de connaître la *distribution* de la population.

III - 1 LOI DES GRANDS NOMBRES

Théorèmes

❶ *Si $\{X_n\}$ une suite de variables aléatoires indépendantes et de même loi, dont les deux premiers moments (l'espérance μ et la variance σ^2) sont finis, la moyenne* $\overline{X} = \frac{1}{n}\sum_i X_i$ *converge en probabilité* vers μ *lorsque* $n \to \infty$.

❷ *Si $\{X_n\}$ une suite de variables aléatoires indépendantes ou non, de lois quelconques, de même espérance μ et telle que :*

$V(\overline{X}_n) \to 0$ *lorsque* $n \to \infty$, alors $\overline{X}_n$ *converge en probabilité vers* μ.

III - 2 THEOREME CENTRAL-LIMITE

Théorème central-limite (ou de la limite centrale)

Soit $\{X_n\}$ une suite de variables aléatoires indépendantes, de même loi de probabilité, d'espérance $\mu = \sum_1^n \mu_i$ *et de variance* $\sigma^2 = \sum_1^n \sigma^2_i$. *On désigne par* $S_n = \sum_1^n X_i$ *et soit la variable centrée réduite associée*

$T_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$, T_n ***converge en loi vers la loi normale*** *centrée réduite* ***(loi standard) N(0,1)*** *quand* $n \to \infty$.

IV - FONCTIONS GENERATRICES

IV - 1 Définition

On appelle ***fonction génératrice*** de la variable X « notée ***(f.g.)*** » la fonction $\psi(t) = E\left[e^{tX}\right]$

➢ ***cas discret*** :

$$\psi(t) = E\left[e^{tX}\right] = \sum_i e^{tx_i} p_i$$

➢ ***cas continu*** :

$$\psi(t) = E\left[e^{tX}\right] = \int_{-\infty}^{+\infty} e^{tx} f(x)dx$$ *t est un paramètre réel.*

Remarques

☞ Cette fonction n'existe pas nécessairement.

☞ Pour assurer la convergence de la série ou de l'intégrale, on considère la fonction à valeurs complexes appelée ***fonction caractéristique*** « notée ***(f.c.)*** » ou ***caractéristique*** de la variable X définie par :

$$\varphi(t) = E\left[e^{itX}\right] = \psi(it) \quad \text{avec} \quad i^2 = -1$$

☞ $\varphi(t)$ existe toujours. C'est la *fonction caractéristique* d'une *fonction de répartition F(x)*. Il *n'existe pas* de *fonction de répartition* autre que *F(x)* dont $\varphi(t)$ soit également la fonction caractéristique.

IV - 2 Propriétés

❶ On peut définir une loi de probabilité par sa *fonction caractéristique*. Pour une ***variable aléatoire continue, la densité de probabilité*** est donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \varphi(t) e^{-itx} dt$$

avec comme *condition,* la *convergence de l'intégrale*.

❷ En utilisant le développement en série entière de e^{itx}

$e^{itx} = 1 + \frac{itx}{1!} + \frac{i^2t^2x^2}{2!} + \ldots + \frac{i^n t^n x^n}{n!} + \ldots$ en sommant puis en intégrant terme à terme et en supposant que tous les moments α_n existent, on obtient :

$$\varphi(t) = 1 + \alpha_1 \frac{it}{1!} + \alpha_2 \frac{i^2t^2}{2!} + \ldots + \alpha_n \frac{i^n t^n}{n!} + \ldots$$

$\varphi(t)$ est une ***fonction génératrice de moments***.

❸ On a encore :

$$\boxed{\frac{d^n\varphi(t)}{dt^n} = i^n\alpha_n \quad \text{pour } t=0}$$

❹ Comme les variables complexes n'ont pas été *introduites*, on **utilise** généralement la **fonction génératrice (*f.g.*)** qui a des propriétés analogues à la *fonction caractéristique* d'où, en supposant que tous les moments α_n existent, on a :

$$\psi(t) = \sum_0^\infty \alpha_n \frac{t^n}{n!} \qquad \frac{d^n\psi(t)}{dt^n} = \alpha_n \quad \text{pour } t=0$$

❺ Lorsque la série précédente est *convergente* pour $t \neq 0$, la *fonction génératrice détermine* complètement la *loi de probabilité*. Dans ce cas, il existe une *correspondance biunivoque* entre une *fonction de répartition* et sa *fonction génératrice*.

IV - 3 Applications

Il s'agit d'énoncer *deux théorèmes importants* dans le calcul des fonctions génératrices, puis de *donner les fonctions génératrices* des lois usuelles.

IV - 3. 1 Théorèmes

Théorème 1

La fonction génératrice d'une somme de variables aléatoires indépendantes est égale au produit des fonctions génératrices des variables composantes.

Théorème 2

*Pour que deux variables aléatoires **X** et **Y** aient la même loi, il est nécessaire et suffisant qu'elles admettent la même fonction génératrice.*

IV - 3. 2 Fonctions génératrices

①- *Loi de Bernoulli : $X \rightarrow$ **B(1, p)***

$$\psi(t)=E\left[e^{tX}\right]=\sum_i e^{tx_i}\text{Prob}(X_i=x_i)$$

d'où $\boxed{\psi(t) = q + p\,e^t}$

②- *Loi Binomiale : $X \rightarrow$ **B(n, p)***

$X = \sum_1^n X_i$. Comme X_i ***suit la loi B(1, p)*** et comme les X_i *sont des variables aléatoires indépendantes*, on a alors :

$$\boxed{\psi(t) = E\left[e^{tX}\right] = (q+pe^t)^n}$$

③- *Loi de Poisson : $X \rightarrow$ **P(λ)***

$$\psi(t) = E\left[e^{tX}\right] = \sum_0^\infty e^{kt}\frac{\lambda^k}{k!}e^{-\lambda} \quad \text{d'où} \quad \boxed{\psi(t) = \exp[\lambda\,(e^t-1)]}$$

④- *Loi normale (loi de Laplace-Gauss) : $X \rightarrow$ **N(m, σ)***

$$\boxed{\psi(t)=E\left[e^{tX}\right]=e^{mt+\frac{1}{2}\sigma^2t^2}}$$

⑤- *Loi gamma : X $\rightarrow$ **Γ(θ, ν)***

$$\boxed{\psi(t)=E\left[e^{tX}\right]=\frac{\theta^\nu}{(\theta\text{-t})^\nu}}$$

V - LOIS DERIVEES

Pour mieux appréhender les phénomènes observés, on cherche à les représenter par des *modèles* basés sur des *fondements* rigoureux et fiables. A partir de ces modèles construits sur des séries statistiques, on peut *analyser, prévoir* et *décider.*

Dans le ***chapitre 7*** du manuel ***« Statistique & Probabilités »***, nous avons introduit des ***modèles discrets*** et *des* ***modèles continus***. Il s'agit de les *compléter* par ceux qui sont les plus fréquemment utilisés dont certains découlent de la *loi normale*.

V - 1 GENERALISATION DE LA LOI UNIFORME

Loi *U(a, b)*

V- 1. 1 Définition

Soit X une variable aléatoire continue uniforme sur $[a, b]$, on considère la variable aléatoire Y définie par :

$$\boxed{Y = (b - a)X + a}$$

Y est une variable aléatoire qui suit la loi $U(a, b)$. En effet la fonction de répartition de Y, $G(y)$ est définie par :

$$G(y) = Prob\,(Y \leq y) \quad d'où \quad G(y) = Prob[(b - a)X + a \leq y]$$

$$= \text{Prob}(X \leq \frac{y-a}{b-a}) = F(\frac{y-a}{b-a}) \quad \text{d'où}$$

$$G(y) = \begin{cases} 0 & \text{si} \quad y < a \\ \dfrac{y-a}{b-a} & \text{si} \quad a \leq y \leq b \\ 1 & \text{si} \quad y > b \end{cases}$$

On en déduit la *densité de probabilité g(y)* :

$$g(y) = G'(y) = \begin{cases} 0 & \text{si } y < a \text{ ou } y > b \\ \dfrac{1}{b-a} & \text{si} \quad a \leq y \leq b \end{cases}$$

V - 1. 2 Caractéristiques de la loi

En utilisant les propriétés de l'espérance et de la variance, on a :

❶ Espérance mathématique

$$E(Y) = E[(b - a)X + a] = (b - a)E(X) + a \quad \text{or} \quad E(X) = \frac{1}{2}$$

$$\text{d'où} \quad \boxed{E(Y) = \frac{b+a}{2}}$$

❷ **Variance**

$V(Y)=V[(b-a)X+a]=(b-a)^2V(X)$

or $V(X)=\frac{1}{12}$ d'où $\boxed{V(X)=\frac{(b-a)^2}{12}}$

V - 1. 3 Utilisation

Recouvrant la *notion d'équiprobabilité*, la loi uniforme se trouve à l'origine de *multiples modélisations* dans des domaines aussi variés que l'économie, la sociologie, la physique, la biologie...

V - 2 LOI GAMMA OU LOI D'ERLANG : γ_ν

V - 2. 1 Définition :

On dit qu'une variable continue X suit la loi d'Erlang ou loi γ_ν de paramètre $\nu \geq 0$ (ν appelé ***degré de liberté***) si *sa densité de probabilité* est donnée par :

$$f(x) = \begin{cases} 0 & si \quad x < 0 \\ \frac{1}{\Gamma(\nu)} e^{-x} x^{\nu-1} & si \quad x \geq 0 \end{cases}$$

avec $\nu \geq 0$ *et* $\Gamma(\nu) = \int_0^{+\infty} e^{-x} x^{\nu-1} dx$ *ayant pour propriété :*

$$\boxed{\Gamma(\nu) = (\nu-1)\,\Gamma(\nu-1)}$$

V - 2. 2 Propriétés de *f(x)*

① $f(x) \geq 0$

② $\int_0^{+\infty} f(x)dx = 1$

③ limite de $f(x) = 0$ lorsque $x \to +\infty$

V - 2. 3 Caractéristiques de la loi γ_ν

❶ **Mode**

C'est la valeur positive de la variable qui rend $f(x)$ maximum.

$Mo = \nu - 1$ *si* $\nu > 1$

$Mo = 0$ *si* $\nu = 1$ *ou* $0 < \nu < 1$

❷ **Espérance mathématique**

On sait que l'espérance est définie par : $E(X) = \int_0^{+\infty} xf(x)dx$

on démontre que : $E(X) = \nu$

❸ **Variance :**

$V(X) = \int_0^{+\infty} x^2 f(x)dx - [E(X)]^2$

on démontre que : $V(X) = \nu$

Remarques:

☞ *La variance* comme *l'espérance* est égale au *paramètre de la loi*

☞ Soient X et Y deux variables *aléatoires indépendantes* suivant respectivement $\gamma(p)$ et $\gamma(q)$ alors la *variable aléatoire* $Z = X + Y$ suit la loi γ $(p + q)$; On dit que *la loi γ est stable.*

V - 2. 4 Cas particuliers :

❶ $\nu = k$ k entier positif $\Gamma(k) = (k-1)!$

a) la *fonction densité de probabilité*

$$f(x) = \frac{1}{(k-1)!} e^{-x} x^{k-1} \quad x > 0$$

b) la *fonction de répartition*

$$F(x) = Prob(X < x) = \int_0^x f(x)dx$$

❷ $\nu = 1$ c'est la loi γ_1 avec $f(x) = e^{-x}$ et $F(x) = 1 - e^{-x}$

γ_1 est appelé **la loi exponentielle** de *paramètre 1*

V- 2. 5 Généralisation de la loi : γ_ν *(a, b)*

Soit X une variable aléatoire qui suit la loi γ_ν Considérons la variable Y définie par

$$Y = a + bX .$$

Cette variable *Y suit la loi γ_ν (a, b)*, son intervalle de définition est $[a, +\infty[$, sa *densité de probabilité* est définie par :

$$g(y) = f(\frac{y-a}{b})\frac{1}{b} = \frac{1}{b\Gamma(\nu)} e^{-\frac{y-a}{b}} (\frac{y-a}{b})^{\nu-1} \quad \text{pour } y > a$$

Ses *caractéristiques* sont données par :

$$E(Y) = E(a + b\,X) = a + b\,E(X) = a + b\nu$$
$$V(Y) = V(a + b\,X) = b^2\,V(X) = b^2\nu \quad \Rightarrow \sigma = b\sqrt{\nu}$$

♦ ***Cas particuliers*** *: $\gamma_1(0, b)$*

$$g(y) = \frac{1}{b} e^{-\frac{y}{b}} = \lambda e^{-\lambda y} \qquad avec \quad y > 0$$

Cette loi est appelée la loi $\xi(\lambda)$ exponentielle de paramètre λ, avec $\lambda = \frac{1}{b}$

Remarques

☞ C'est une généralisation de la loi γ_1

☞ $\xi(\lambda)$ est identique à la loi $\gamma_1(0, \frac{1}{\lambda})$

☞ Elle est *définie* pour tout *x positif* par :

① la *densité de probabilité* : $f(x) = \lambda e^{-\lambda x}$

② la *fonction de répartition* : $F(x) = 1 - e^{-\lambda x}$

③ les *caractéristiques* de la loi sont : $E(X) = \frac{1}{\lambda}$ et $V(X) = \frac{1}{\lambda^2}$

V - 2. 6 Utilisation

La *famille de loi* γ_ν est utilisée en *économie* (dans l'étude des amortissements des investissements, ...) mais en *médecine* également.

D'une manière générale, elle est employée lorsqu'on étudie des *modèles de durée de vie* c'est à dire des événements aléatoires en fonction du temps (processus de Poisson).

V - 3 LOI LOG-NORMALE

V - 3. 1 Définition

On dit qu'une variable aléatoire continue X suit une loi *log-normale* (*logarithmo-normale*) si son *logarithme suit une loi normale*.

On pose $Y = ln\,X$

✓ Si X suit une loi *log-normale*, Y suit une *loi normale*.

✓ Si μ est la moyenne de Y, σ son écart-type et si T est une loi normale centrée réduite $N(0, 1)$, on a :

$Y = \mu + \sigma T$ et $X = e^{\mu+\sigma T}$ les valeurs de X sont des nombres positifs appartenant à l'intervalle $[0, +\infty[$

① ***Densité de probabilité***

$$f(x) = \begin{cases} 0 & si \quad x < 0 \\ \frac{1}{\sigma x\sqrt{2\pi}} \exp \frac{(\ln x - m)^2}{2\sigma^2} & si \quad x \geq 0 \end{cases}$$

② ***Fonction de répartition***

$$F(x) = \text{Prob}(X < x) = \int_0^x f(x)dx$$

V - 3. 2 Caractéristiques

❶ **Mode** $Mo = exp(m - \sigma^2)$

❷ **Espérance mathématique**

$$E(X) = exp(m + \frac{\sigma^2}{2})$$

❸ **Variance**

$$V(X) = (e^{\sigma^2} - 1)e^{2m+\sigma^2}.$$

V - 3. 3 Utilisation

Cette loi est utilisée à chaque fois qu'un processus est provoqué par un *grand nombre de causes* qui agissent *multiplicativement*. C'est le cas des communications téléphoniques, des distributions de revenus, ...

V - 4 LOI DU KHI-DEUX : χ^2

V - 4 .1 Définition

Soient $X_1, \ldots, X_i, \ldots, X_n$ une suite de n *variables aléatoires normales centrées réduites et indépendantes.*

La variable aléatoire $\chi_n^2 = X_1^2 + \ldots X_i^2 + \ldots X_n^2 = Z_n$ suit une loi de *probabilité de densité* :

$$f(z) = \begin{cases} 0 & si \quad z \leq 0 \\ \frac{1}{2}\frac{1}{\Gamma(\frac{n}{2})}(\frac{z}{2})^{\frac{n}{2}-1} e^{-\frac{z}{2}} & si \quad z > 0 \end{cases}$$

La variable aléatoire χ^2 est dite variable ***Khi-deux*** « ***Khi-carré*** » à *n degrés de liberté* ***(d. d. l.)***

V - 4 . 2 <u>Caractéristiques</u>

La loi de χ^2 est la loi $\gamma(\frac{n}{2},\frac{1}{2})$ où n est un entier positif, on peut alors déterminer ses caractéristiques à partir de celles de la loi gamma γ

❶ **Espérance mathématique** $E(\chi^2_n) = n$

❷ **Variance** $V(\chi^2_n) = 2n$

❸ **Stabilité**

✓ La *somme* de *deux lois* de χ^2 est *une loi* de χ^2

✓ Si X et Y sont deux *variables aléatoires indépendantes* telles que

$X \rightarrow \chi^2(n_1)$ et $Y \rightarrow \chi^2(n_2)$ alors $X + Y \rightarrow \chi^2(n_1 + n_2)$

V - 4 .3 <u>Approximation de la loi</u>

✓ *Lorsque n* tend vers *l'infini*, la valeur de la *variable aléatoire*

$$T = \frac{\chi^2_n - n}{\sqrt{2n}} \rightarrow N(0,1)$$

✓ *En pratique* si $n > 30$ l'expression

$$\sqrt{2\chi^2_n} - \sqrt{2n-1} \rightarrow N(0,1)$$

V - 4. 4 Table de χ^2

Pour les *calculs* relatifs à *la loi* de χ^2, il existe une *table de fractiles*. C'est une table à double entrée. Les *valeurs de* χ^2_α, pour différentes valeurs du *degré de liberté* ν et de la *probabilité* α, sont données par la *table II* en annexes

✧ La *surface* limitée par la *courbe* et *l'axe des abscisses* vaut :

$$\int_0^{+\infty} f(\chi^2) d\chi^2 = 1$$

✧ *La surface correspond à* α *tel que* :

$$\alpha = 1 - \int_0^{\chi_\alpha^2} f(\chi^2) d\chi^2 = \text{Prob}(\chi^2 \geq \chi_\alpha^2)$$

✧ *La valeur de χ^2 est lue en fonction de α et du nombre de degré de liberté*

$$\nu = n$$

✧ Prob*(a < x < b) = F(b) - F(a) permet de déterminer des probabilités d'intervalles lorsqu'on dispose d'une table détaillée. Sinon il faut recourir à des interpolations linéaires.*

V - 4. 5 Utilisation

Dans le manuel *« **Statistique & Probabilités** »* nous avons signalé que la loi du ***Khi-deux*** sert *à juger* de *la validité de l'ajustement d'une distribution observée à une distribution théorique*. Nous y reviendrons dans le ***chapitre 6*** consacré à *l'utilisation de χ^2*.

V - 5 LOI DE STUDENT

V - 5. 1 Définition

Soit *(X, Y)* un *couple de variables aléatoires indépendantes* telles que $X \rightarrow N(0, 1)$ et $Y \rightarrow \chi^2(n)$ alors la variable aléatoire

$U = \dfrac{X}{\sqrt{Y}}$ suit une loi *de densité de probabilité*

$$h(u) = \frac{1}{\beta(\frac{1}{2}, \frac{n}{2})} \frac{1}{(1+u^2)^{\frac{n+1}{2}}} \quad \text{avec } u \in R$$

La variable aléatoire $t = \sqrt{n}U = \dfrac{X}{\sqrt{\frac{Y}{n}}}$ suit une loi de *densité de probabilité*

$$h(t) = \frac{1}{\sqrt{n}} h(\frac{t}{\sqrt{n}}) = \frac{1}{\sqrt{n}\beta(\frac{1}{2}, \frac{n}{2})} \times \frac{1}{(1+\frac{t}{n}^2)^{\frac{n+1}{2}}} \text{ avec } t \in R$$

avec $\beta(p,q) = \dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 t^{p-1}(1-t)^{q-1} dt$ (p et q positifs)

La *loi de probabilité* de t est appelée ***loi de Student*** à ***n degrés de liberté***.

♦ ***Cas particulier***

La loi de *Student à 1 degré de liberté* porte le nom de **loi de Cauchy**

V - 5. 2 Caractéristiques de la loi

On admettra que :

pour $n > 1$ $E(t) = 0$

pour $n > 2$ $\boxed{V(t) = \frac{n}{n-2}}$

***Pour n** = 1 et n =2 la loi n'admet pas de variance finie.*

V - 5. 3 Approximation de la loi

On démontre que lorsque ***n*** tend vers *l'infini* la loi limite de ***la loi de Student*** est la *loi standard N(0, 1)*

V - 5. 4 Table de Student t_ν

✧ Pour les calculs relatifs à la loi de Student t_ν, il existe une table des fractiles. La *table III* donne les valeurs de t_α pour différentes valeurs du degré de liberté ν et de la probabilité α.

✧ *Les valeurs de* t *sont liées aux valeurs de* α *telles que :*

$$\boxed{\text{Prob}(t \geq t_{\alpha,\nu}) = \alpha}$$

$t_{\alpha,\nu}$ *est une valeur fixée qui se lit sur la table.*

✧ *La distribution de Student* t_ν *est une famille de distribution car pour chaque valeur de* ν*, il existe une distribution*

✧ *A cause de la symétrie*

$$\boxed{t_{\alpha,\nu} = -\, t_{(1-\alpha),\nu}}$$

V - 5. 5 <u>Utilisation</u>

Cette loi s'applique à l'étude de la moyenne d'une *loi de Gauss* dont la *variance* est *inconnue* et lorsque la *taille de l'échantillon* est *petite*. Elle est utilisée dans la *théorie de l'estimation et celle des tests.*

V - 6 <u>LOI DE FISHER-SNEDECOR</u>

V - 6. 1 <u>Définition</u>

Soit (X, Y) un couple de variables aléatoires indépendantes telles que :

$X \rightarrow \chi^2(n_1)$ et $Y \rightarrow \chi^2(n_2)$ alors la variable aléatoire

$U = \sqrt{\frac{X}{Y}}$ suit une loi de probabilité de densité

$$h(u) = \begin{cases} 0 & u \leq 0 \\ \dfrac{2}{\beta(\frac{n_1}{2}, \frac{n_2}{2})} \dfrac{u^{n_1-1}}{(1+u^2)^{\frac{n_1+n_2}{2}}} & u > 0 \end{cases}$$

On dit que U suit une ***loi de Fisher-Snedecor à*** $(n_1 + n_2)$ *degrés de liberté*

$$\beta(\frac{n_1}{2}, \frac{n_2}{2}) = \frac{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}{\Gamma(\frac{n_1+n_2}{2})}$$

V - 6. 2 <u>Caractéristiques de la loi</u>

❶ Espérance mathématique

$$E(Z) = \frac{n_2}{n_2 - 2} \quad \text{si} \quad n_2 > 2$$

❷ Variance

$$V(Z) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)} \quad si \quad n_2 > 4$$

Remarques

☞ $Z = U^2$ suit une loi de densité de probabilité : $g(z) = \frac{1}{2\sqrt{v}} h(u)$

☞ $F = \frac{\frac{X}{n_1}}{\frac{Y}{n_2}} = \frac{n_2}{n_1} Z$ F suit la ***loi de Fisher-Snedecor*** de densité :

$$\begin{cases} 0 & f \le 0 \\ \dfrac{(\frac{n_1}{n_2})^{\frac{n_1}{2}} f^{\frac{n_1}{2}-1}}{\beta(\frac{n_1}{2},\frac{n_2}{2})(1+\frac{n_1}{n_2}f)^{\frac{n_1+n_2}{2}}} & f > 0 \end{cases}$$

☞ $Z = \ln\sqrt{F} \Leftrightarrow F = e^{2z}$ z est appelée ***la variable de Fisher***. Sa densité de probabilité s'obtient à partir de la densité de probabilité de F par un changement de variable.

V - 6. 3 Table de Fisher-Snedecor

✧ *Cette loi est tabulée. La table donne les valeurs de F ayant la probabilité d'être dépassée* $\text{Prob}(X > x_\alpha) = \alpha$ *avec* $X \rightarrow F(n_1, n_2)$ *et* α *prenant uniquement deux valeurs* $\alpha = 0,05$ *et* $\alpha = 0,01$

✧ *On en déduit également*

$$\boxed{x_{1-\alpha}(n_2, n_1) = \frac{1}{x_\alpha(n_2, n_1)}}$$

V - 6. 4 Utilisation

Cette loi est utilisée lorsqu'on *compare deux populations* par rapport à une *certaine caractéristique* (moyenne, variance..). Son rôle est essentiel dans les méthodes d'analyse de la variance.

V - 7 LOI DE PARETO

V - 7. 1 Définition

C'est une loi de paramètre $\alpha > 0$ dont la densité de probabilité est définie par :

$$f(x) = \begin{cases} 0 & \text{si } x < x_0 \\ \dfrac{\alpha x_0^\alpha}{x^{\alpha+1}} & \text{avec } x \ge x_0 > 0 \end{cases}$$

V - 7. 2 Caractéristiques de la loi

❶ Espérance mathématique

$$E(X) = \frac{\alpha x_0}{\alpha - 1}$$

❷ Variance

$$V(X) = \frac{\alpha x_0^2}{(\alpha - 1)^2(\alpha - 2)} \qquad \text{pour } \alpha > 2$$

V - 7. 3 Utilisation

Cette loi est utilisée en théorie d'assurances et pour la modélisation des distributions de revenus. Sa fonction *densité de probabilité* peut s'interpréter comme le *revenu minimum* en fonction d'un *paramètre* $\alpha > 0$ en général *voisin de 2*.

CONCLUSION GENERALE

▪ Nous nous sommes limités dans ce chapitre *à rappeler les modes de convergence* et *à compléter les lois que nous avons jugées les plus utilisées dans des situations concrètes*. Cependant d'autres *modèles continus* peuvent présenter un intérêt pour la modélisation, citons à titre indicatif : la loi *Béta*, la loi *logistique*, la loi de *Weibull*, la loi *binomiale négative*...

▪ La *loi de Gauss demeure essentielle pour l'approximation des lois* discrètes ou des lois continues. Son rôle demeure déterminant dans l'inférence statistique.

▪ D'autres lois usuelles ont été déjà étudiées dans le manuel « **Statistique & Probabilités** »

Auto-évaluation ?

Répondre par *Vrai* ou *Faux* :

1- La variable aléatoire de Student est une v. a. discrète.
2- La distribution de Student est symétrique par rapport à zéro.
3- Le nombre de degrés de liberté pour la distribution de Student est égal à la taille de l'échantillon.
4- La loi γ a un seul mode.
5- La loi exponentielle est un cas particulier de la loi γ

6- Indiquer ***les affirmations correctes :***

① Lorsqu'un processus est provoqué par un grand nombre de causes qui agissent additivement, on utilise :
a) la loi normale b) la loi log-normale.

② Lorsqu'un processus est provoqué par un grand nombre de causes qui agissent multiplicativement, on utilise :
a) la loi normale b) la loi log-normale.

③ Si X suit une loi exponentielle de paramètre $\lambda = 0$ et $\lambda = 3$
a) l'espérance de X est égale à sa variance
b) la fonction de répartition de X est une fonction continue
c) X ne peut prendre que des valeurs supérieures à 3.

④ Soit Y une somme de n variables aléatoires indépendantes :
a) Si les X_i suivent des lois normales, alors Y suit une loi normale
b) Si les X_i suivent des lois exponentielles, alors Y suit une loi exponentielle.

7- ***Utilisation des tables***

a) Distribution Normale :

En supposant que le quotient intellectuel (QI) d'un groupe

d'individus suit la distribution normale N(100, 10), calculer :

① La proportion des individus ayant un QI compris entre 100 et 105

② La proportion des individus ayant un QI inférieur à 80

③ La proportion des individus ayant un QI supérieur à 120

④ Le QI dépassé par 10% des individus.

b) Distribution de la loi Khi-deux χ^2

① Calculer les probabilités suivantes :

$\text{Prob}(\chi_1^2 \geq 3{,}841)$ $\quad$ $\text{Prob}(\chi_5^2 > 11{,}07)$

$\text{Prob}(\chi_6^2 \geq 12{,}592)$

② Vérifier que :

$\text{Prob}[(\chi_{20}^2 > 37{,}566)] = 0{,}01$

$\text{Prob}[(\chi_{18}^2 > 22{,}760)] = 0{,}20$

c) Distribution de la loi de Student t_ν

① Calculer les valeurs de t pour les quantités suivantes

*$\alpha = 0{,}01$; $\nu = 12$ $\quad$ *$\alpha = 0{,}025$; $\nu = 20$ $\quad$ *$\alpha = 0{,}10$; $\nu = 24$

*$\alpha = 0{,}75$; $\nu = 10$ $\quad$ *$\alpha = 0{,}0025$; $\nu = 30$ $\quad$ *$\alpha = 0{,}025$; $\nu = 90$

② Si t_ν est une variable de Student à 16 degrés de liberté, calculer t :

* $\text{Prob}[|t| \leq t] = 0{,}95$ $\quad$ * $\text{Prob}[|t| \leq t] = 0{,}99$ $\quad$ * $\text{Prob}[|t| \leq t] = 0{,}999$.

d) Distribution de la loi de Fisher

Les tables de Fisher donnent pour différentes valeurs de degrés de liberté les valeurs de x telles que la fonction de répartition de F prend les valeurs 0,90 ; 0,95 ; 0,975 ; 0,99 ; 0,995.

Pour calculer les valeurs de x telles que *la fonction* de répartition de F prend les valeurs 0,10 ; 0,05 ; 0,025 ; 0,01; 0,005 on procède comme suit :

$$\text{Prob}[F \leq x] = \text{Prob}[\frac{1}{x} \leq \frac{1}{F_{n_1,n_2}}] = \text{Prob}[\frac{1}{x} \leq F_{n_2,n_1}]$$

Déterminer x tel que :

* $F_{7,9}(x) = 0{,}9$ $\quad$ * $F_{7,9}(x) = 0{,}1$ $\quad$ * $F_{7,9}(x) = 0{,}01$

$* F_{8,15}(x) = 0{,}9$ $* F_{20,20}(x) = 0{,}95$ $* F_{15,8}(x) = 0{,}025$

8- Démontrer la relation $\Gamma(\nu) = (\nu - 1)\Gamma(\nu - 1)$

9- Démontrer les propriétés de *f(x)* et calculer ses caractéristiques ?

10- Démontrer que $\Gamma(k) = (k-1)!$

11- La variable aléatoire centrée réduite *T* suit la loi de χ^2 ?

12- Calculer les valeurs critiques de la loi de Khi-deux pour
$n = 10$ et $\alpha = 5\%$; $n = 25$ et $\alpha = 1\%$

13- Calculer $\beta(\frac{1}{2};\frac{1}{2})$

14- Déterminer les valeurs de ***T*** telle que :
$\text{Prob}(|T| < x) = 0{,}7$ avec $n = 8$

15- La variable aléatoire *X* suit la loi de *Fisher-Snedecor F(3; 5)*,
avec $n_1 = 3$ et $n_2 = 5$,
déterminer *x* telle que :
$\text{Prob}(X > x) = 0{,}05$; $\text{Prob}(X > x) = 0{,}01$

16- La variable aléatoire X suit la loi de *Fisher-Snedecor F(5; 3)*,
avec $n_1 = 5$ et $n_2 = 3$.
Déterminer *x* telle que $\text{Prob}(X > x) = 0{,}95$

Exercices ?

Exercice 1

Soient $X_1,\ldots,X_i,\ldots,X_n$, n variables aléatoires réelles indépendantes telles que $X \to B(n;p)$, montrer que X_n converge en distribution de probabilité vers la *loi de Poisson* de paramètre $\lambda = np$.

D'après examen janvier 1998, ISG Tunis.

Exercice 2

Soit X une variable aléatoire qui suit la loi uniforme sur *[a, b]*

1 - Calculer l'espérance et la variance de X.

2 - Soient $X_1, \ldots, X_i, \ldots, X_n$, n variables aléatoires réelles indépendantes de même loi que X. Déterminer n pour que :

$$\text{Prob}\left(\left|\overline{X} - \frac{a+b}{2}\right| < 10^{-3}\right) \leq 0{,}95 \text{ en utilisant :}$$

① l'inégalité de Bienaymé-Tchebychev

② le théorème central-limite.

Application numérique : $a = 0$ et $b = 1$

3 - On suppose que :

$$a = -\frac{\theta}{2} \qquad b = \frac{\theta}{2} \text{ avec } \theta > 0 \qquad S_n = \sum_i X_i \text{ et } \overline{X_n} = \frac{S_n}{n}$$

① Déterminer, pour n «assez grand », les lois asymptotiques de S_n et $\overline{X_n}$

② On prend $\theta = 10^{-n}$ et $n = 10^6$, calculer $\text{Prob}\left(|S_n| < \frac{1}{2}10^{-n+3}\right)$

③ Montrer que $\overline{X_n}$ converge en probabilité vers une variable à déterminer.

D'après examen février 1995-1996, ISG Tunis.

Exercice 3

Soient $X_1,\ldots,X_i,\ldots,X_n$, n variables aléatoires réelles indépendantes et telles que : $X_i \to N(m_i;\sigma_i^2)$ avec $i = 1,\ldots,i,\ldots,n$

1- Donner l'expression de la fonction génératrice de X_i

2- Soit $Y = \sum_1^n a_i X_i$, une combinaison linéaire des X_i

Déterminer la fonction génératrice de Y. En déduire la loi de Y.

D'après examen 1996-1997, ISG Tunis.

Exercice 4

Soient X_1 et X_2 deux variables aléatoires indépendantes. On suppose que X_1 et $Y = X_1 + X_2$ suivent respectivement les lois $\chi^2(r_1)$ et $\chi^2(r)$ avec $r_1 < r$

Montrer que X_2 suit la loi $\chi^2(r - r_1)$

D'après examen 1996-1997, ISG Tunis.

Exercice 5

Soient $(X_i)_{i=1}^r$ n variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de même loi de paramètre λ et soient $Y_n = \frac{1}{n}\sum_1^n X_i$.

Montrer que Y_n converge en probabilité vers λ en utilisant :

1- l'inégalité de Bienaymé-Tchebychev

2- la fonction génératrice des moments.

D'après examen 1996-1997, ISG Tunis.

Exercice 6

Soit X une v. a. à valeurs dans N. On pose $p_k = \text{Prob}(X = k)$. On considère l'application : $G_X : [0,1] \to R; \quad t \to E(t^x)$

On désigne par G' et G'' les premières dérivées de G :

1- Calculer G' et G'' ; en déduire que : $E(X) = G'$; $E[X(X-1)] = G''$

2- On pose $p_k = \text{Prob}(X = k) = e^{-m}\frac{m^k}{k!}$ avec $k \in N$ et $m > 0$.

Déterminer $G_X(t)$. En déduire $E(X)$ et $V(X)$.

3- Soient X et Y deux variables indépendantes suivant respectivement une *loi de Poisson* de paramètre a et b.

On pose $S = X + Y$. Déterminer $G_S(t)$.

Corrigés

Auto-évaluation

1- Vrai
2- Vrai
3- Faux
4- Faux
5- Vrai

6-
1- a
2- b
3- c
4- a ; b

7- ***Utilisation des tables***

a) Distribution de la loi Normale

① Prob*(100 ≤ X ≤ 105) = F(105) - F(100) = G(0,6913) - G(0)*
= 0,1915 - 0,5 = 0,1915

② Prob*(X < 80)= 0, 0228*

③ Prob*(X ≥ 120) = 1* - Prob*(Y ≤ 120) = 0,022*

④ Prob*(X ≥ QI) = 0,1= 1* - Prob*(X ≤ QI)* $\Rightarrow$ *0,9 = G(QI - 10)*
$\Rightarrow$ *QI = 11,29*

b) Distribution de la loi Khi-deux χ^2

1) Prob$(\chi_1^2 \geq 3{,}841) = 0{,}05$; Prob$(\chi_5^2 > 11{,}07) = 0{,}05$;
Prob$(\chi_6^2 \geq 12{,}592) = 0{,}05$

2) On vérifie que :

$$\text{Prob}[(\chi^2_{20} > 37{,}566)] = 1 - 0{,}99 = 0{,}01 \ ;$$

$$\text{Prob}[(\chi^2_{18} > 22{,}760)] = 1 - 0{,}8 = 0{,}20$$

c) Distribution de la loi de Student t_ν

① Calcul des valeurs de t

* t = 2,6810 * t = 2,0860 *t = 1,3178 * t = 0,6998

En utilisant l'approximation de la loi normale, on trouve :

*t = 2,0423 * t = 1,9867

② Si t_ν est une variable de Student à 16 degrés de liberté

$\nu = 16$, calcul de *t :1,746 *t = 3,686 *t = 4,015

d) Distribution de la loi de Fisher

Les tables de Fisher donnent, pour différentes valeurs de degrés de liberté, les valeurs de x telles que la fonction de répartition de F prend les valeurs 0,90 ; 0,95 ; 0,975 ; 0,99 ; 0,995.

Pour calculer les valeurs de x telles que la fonction de répartition de F prend les valeurs 0,10 ; 0,05 ; 0,025 ; 0,01; 0,005 ; on procède comme suit :

$$\text{Prob}[F \le x] = P[\frac{1}{x} \le \frac{1}{F_{n_1, n_2}}] = P[\frac{1}{x} \le F_{n_2, n_1}]$$

Calcul de x :

*x = 2,51 *x = 0,368 *x = 0,149

*x = 3,2 *x = 2,12 *x = 0,313

8- Calcul de $\Gamma(\nu) = (\nu - 1)\Gamma(\nu - 1)$

On sait que : $\Gamma(\nu) = \int_0^{+\infty} e^{-x} x^{\nu-1} dx$

on pose alors : $e^{-x} dx = du$ et $x^{\nu-1} = v$. En intégrant par partie, on obtient :

$$\Gamma(\nu) = [-e^{-x} x^{\nu-1}]_0^{+\infty} + (\nu - 1)\int_0^{+\infty} e^{-x} x^{\nu-2} dx$$

d'où la relation : $\Gamma(\nu) = (\nu - 1)\Gamma(\nu - 1)$

9- Propriétés de $f(x)$ et calcul de ses caractéristiques :

$* \int_0^{+\infty} f(x)dx = \frac{1}{\Gamma(\nu)}\int_0^{+\infty} e^{-x}x^{\nu-1}dx = \frac{\Gamma(\nu)}{\Gamma(\nu)} = 1$

$* f(x) \geq 0$ car $\Gamma(\nu) > 0$; $e^{-x} > 0$ et $x \geq 0$

$* \lim_{x\to+\infty} f(x)dx = \frac{1}{\Gamma(\nu)}\lim_{x\to+\infty}(e^{-x}x^{\nu-1}) = 0$

*Calcul du mode :

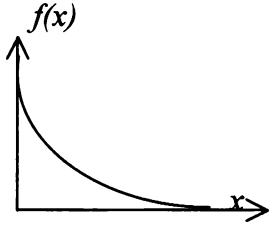
On calcule la dérivée $f'(x) = \frac{1}{\Gamma(\nu)}[-e^{-x}x^{\nu-1}] + (\nu-1)e^{-x}x^{\nu-2}]$

$$f'(x) = \frac{1}{\Gamma(\nu)}e^{-x}x^{\nu-2}[(\nu-1)-x]$$

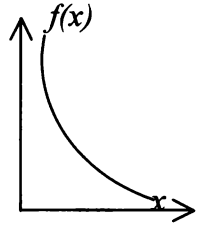
d'où $f'(x) = 0$ pour $x = \nu - 1$ *or* $x \geq 0$ *soit* $\nu - 1 \geq 0$

Si $* \nu - 1 = 0 \Rightarrow \nu = 1$ d'où $Mo = 0$

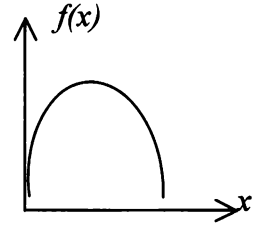
$* \nu > 1 \Rightarrow$ d'où $Mo = \nu - 1$



$\nu = 1$



$0 < \nu < 1$



$\nu > 1$

$$E(X) = \int_0^{+\infty} f(x)dx = \frac{1}{\Gamma(\nu)}\int_0^{+\infty} e^{-x}x^{\nu}dx = \frac{\Gamma(\nu+1)}{\Gamma(\nu)} = \nu$$

$$V(X) = \int_0^{+\infty} x^2 f(x)dx - \nu^2 = \frac{\Gamma(\nu+2)}{\Gamma(\nu)} - \nu^2 = \nu$$

10- Calcul de $\Gamma(k) = (k-1)!$

$$\Gamma(k) = (k-1)!\Gamma(1)$$

or $\Gamma(1) = \int_0^{+\infty} e^{-x} dx = [e^{-x}]_0^{+\infty} = 1$ d'où la relation cherchée.

11- $T = \dfrac{x-m}{\sigma}$ *suit une loi de* χ^2

On pose $Y = T^2$, sa fonction de répartition est définie, pour $y > 0$, par

$$\text{Prob}(Y < y) = \text{Prob}(-\sqrt{y} < T < \sqrt{y}) = G(\sqrt{y}) - G(-\sqrt{y})$$

et sa densité de probabilité par

$$g(y) = \frac{e^{-y/2}}{\sqrt{2\pi y}} \text{ cette expression s'écrit : } g(y) = \frac{1}{2}\frac{e^{-y/2}\, y^{-1/2}}{\Gamma(1/2)}$$

c'est à dire la densité de probabilité de la loi de $\chi^2(1)$.

12- On lit sur la table les valeurs correspondantes de χ^2_α

$n = 10$; $\alpha = 5\%$ $\Rightarrow \chi^2 = 18,31$

$n = 25$; $\alpha = 1\%$ $\Rightarrow \chi^2 = 44,31$

On a une chance sur *100* d'observer la valeur de *Khi-deux* supérieure à *18,31* pour le premier cas et à *44,31* pour le second cas.

13- Calcul de $\beta(\frac{1}{2};\frac{1}{2})$

On a $\beta(\frac{1}{2};\frac{1}{2}) = \dfrac{\Gamma(1/2)\Gamma(1/2)}{\Gamma(1)} = \int_0^1 \dfrac{dt}{\sqrt{t(1-t)}} = \pi$

14- On lit sur la table de *Student* les valeurs correspondantes

$n = 8$; $\alpha = 0,7$ $\Rightarrow x = 0,339$

15- On lit sur la table loi de *Fisher-Snedecor* $n = 3$ et $n_2 = 5$

$Prob(X > x) = 0,05$ $\Rightarrow x = 5,41$

$Prob(X > x) = 0,01$ $\Rightarrow x = 12,06$

16- La valeur de $x_{1-\alpha}$ n'existe pas sur la table , on calcule alors

$$\text{Prob}(X \geq \frac{1}{x_\alpha}) = 0,95 \text{ avec } n_1 = 5 \text{ et } n_2 = 3$$

d'où $x_{1-\alpha} = \dfrac{1}{x_\alpha(3,5)} = \dfrac{1}{5,41} = 0,1848$

Exercices

Exercice 1

On sait que $M_{X_i(t)} = (1 - p + p\,e^t)^n = (1 - \frac{\lambda}{n} + \frac{\lambda}{n} e^t)^n$

Cette expression s'écrit : $[1 + \frac{\lambda(e^t - 1)}{n}]^n$

Or

$\lim_{n \to \infty} (1 + \frac{x}{n})^n = e^x$ d'où $\lim_{n \to \infty} [1 + \frac{\lambda(e^t - 1)}{n}]^n = e^{\lambda(e^t - 1)}$

Cette limite est la fonction génératrice de la loi de Poisson.

Exercice 2

1- *Calcul de l'espérance et de la variance de X*

On sait que :

$$E(X) = \frac{a+b}{2} \quad \text{et} \quad V(X) = \frac{(b-a)^2}{12}$$

2- *Détermination de n*

$$V(\overline{X}) = \frac{(b-a)^2}{12n}$$

① En appliquant l'inégalité de *Bienaymé-Tchebychev*, on a :

$$\text{Prob}\left(\left| \overline{X} - \frac{a+b}{2} \right| < 10^{-3} \right) \geq 0{,}95 \Rightarrow 1 - \frac{(b-a)^2}{12n10^{-6}} \# 0{,}95$$

d'où $n = \frac{(b-a)^2 10^8}{12 \times 5}$

Pour $a = 0$ et $b = 1$ on obtient : $n = 1.666.667$

② En appliquant le *théorème central limite* , on obtient :

$$\overline{X} \to N(\frac{a+b}{2}, \frac{(b-a)^2}{12n})$$

$$\text{Prob}(\left| \overline{X} - \frac{a+b}{2} \right| < 10^{-3}\frac{\sqrt{12n}}{b-a}) = 2G(10^{-3}\frac{\sqrt{12n}}{b-a}) - 1$$

$$\Rightarrow 10^{-3}\ \frac{\sqrt{12n}}{b-a} \# 1{,}96$$

Application numérique :

Pour $a = 0$ et $b = 1$ on obtient : $n = 163.333$

3-

① Détermination de la loi de S_n et $\overline{X}_n$ pour n assez grand :

Avec $a = -\frac{\theta}{2}$ et $b = \frac{\theta}{2}$, on a : $E(X) = 0$ et $V(X) = \frac{\theta^2}{12}$

et par application du *théorème central-limite* :

$$S_n \to N(0; n\frac{\theta^2}{12}) \text{ et } \overline{X}_n \to N(0; \frac{\theta^2}{12n})$$

② Calcul de $\text{Prob}(\left| S_n \right| < \frac{1}{2}10^{-n+3})$ pour $n = 10^6$

$$\text{Prob}(\left| \frac{S_n}{\theta\sqrt{\frac{n}{12}}} \right| < \frac{1}{2}\frac{10^{-n+3}}{\theta\sqrt{\frac{n}{12}}}) = 2G(\frac{1}{2}\frac{10^{-n+3}}{\theta\sqrt{\frac{n}{12}}}) - 1 = 2G(1{,}73) - 1 = 0{,}9164$$

③ $\overline{X}_n$ converge en probabilité vers 0.

Exercice 3

① La fonction génératrice de la loi normale est donnée par :

$$M_{X_i(t)} = e^{m_i t + \frac{1}{2}\sigma_i^2 t^2}$$

② Calcul de la fonction génératrice de Y

$$M_{Y(t)} = M_{\sum_1^n a_i X_i}(t) = \prod_1^n M_{X_i}(a,t)$$

$$M_{Y(t)} = e^{t\sum_1^n a_i m_i + \frac{1}{2}\sum_1^n (a_i^2 \sigma^2{}_i) t^2}$$

d'où $Y \rightarrow N(\sum_1^n a_i m_i ; \sum_1^n a_i{}^2 m_i{}^2)$

Exercice 4

X_2 suit la loi $\chi^2(r - r_1)$

X_1 et X_2 étant deux variables aléatoires indépendantes, on a :

$$X_1 \rightarrow \chi^2(r_1) \qquad \text{et} \qquad M_{X_1(t)} = (1-2t)^{\frac{-r_1}{2}}$$

$$X_2 \rightarrow \chi^2(r_2) \quad \text{et} \quad M_{X_2(t)} = (1-2t)^{\frac{-r_2}{2}}$$

Avec $Y = X_1 + X_2 \rightarrow M_{Y(t)} = (1-2t)^{\frac{-r}{2}}$

Or $M_{Y(t)} = M_{X_1(t)} M_{X_2(t)} \Rightarrow M_{X_2(t)} = \dfrac{M_{Y(t)}}{M_{X_1(t)}}$

Soit : $M_{X_2(t)} = \dfrac{(1-2t)^{\frac{-r}{2}}}{(1-2t)^{\frac{-r_1}{2}}} = (1-2t)^{-\frac{r-r_1}{2}}$

d'où X_2 suit la loi $\chi^2(r - r_1)$.

Exercice 5

$(X_i)_1^r$ n v. a. indépendantes et identiquement distribuées(i. i. d.) de même loi de paramètre λ avec $Y_n = \dfrac{1}{n}\sum_1^n X_i$. Y_n converge en probabilité :

① *Inégalité de Bienaymé-Tchebychev*

$\forall \varepsilon > 0 \quad \text{Prob}(\,| Y_n - \lambda | < \varepsilon) \rightarrow 1$ lorsque $n \rightarrow \infty$

$$E(Y_n) = \lambda \ \forall \varepsilon > 0 \quad \text{Prob}(\mid Y_n - \lambda \mid < \varepsilon) \to 1 - \frac{V(Y_n)}{\varepsilon^2} = 1 - \frac{\lambda}{n\varepsilon^2}$$

qui tend vers 1 lorsque $n \to \infty$

② la fonction génératrice des moments :

$$M_{Y_n}(t) = E(e^{tY_n}) = \prod_1^n M_{X_i}\left(\frac{t}{n}\right)$$

En utilisant le développement de e^t, on obtient :

$$[(e^{-\lambda + \lambda(1 + \frac{t}{n} + \ldots)})]^n = e^{\lambda \frac{t}{n} n} = e^{\lambda t} \Rightarrow Y_n \to \lambda$$

Exercice 6

1 - *Calcul de G' et de G''*

$G_X(t) = E(t^X)$ $\quad G'_X(t) = E(X\, t^{X-1})$; $G''_X(t) = E[X(X-1)\, t^{X-2}]$

$G'_X(1) = E(X)$ $\quad G''_X(1) = E[X(X-1)]$

2 - On pose $p_k = \text{Prob}(X = k) = e^{-m} \dfrac{m^k}{k!}$ avec $k \in N$ et $m > 0$.

$$G_X(t) = E(t^X) = \sum_k t^k e^{-m} \frac{m^k}{k!} = e^{m(t-1)}$$

d'où $G'_X(t) = me^{m(t-1)}$ $\quad G''_X(t) = m^2 e^{m(t-1)}$

et $E(X) = m$; $E[X(X-1)] = m^2 \Rightarrow V(X) = m$

3 - *Détermination de $G_S(t)$*

X et Y deux variables indépendantes suivant respectivement une *loi de Poisson* de paramètre a et b : $S = X + Y$

① $\text{Prob}(S = s) = \sum_0^s P(X = k) P(Y = s - k)$

$$= \sum_0^s e^{-(a+b)} \frac{a^k b^{s-k}}{k!(s-k)!} = \frac{e^{-(a+b)}}{s!} (a+b)^s$$

d'où S *suit une loi de Poisson de paramètre (a + b)*

② En utilisant les *fonctions génératrices* :

$G_S(t) = G_{X+Y}(t) = G_X(t)\, G_Y(t) = \exp[a(e^t - 1)] \exp[b(e^t - 1)]$

$= \exp[(a+b)(e^t - 1)]$

En utilisant *la loi de Poisson* : $G_S(t) = G_{X+Y}(t) = \exp[(a+b)(e^t - 1)]$

Chapitre 2

ECHANTILLONNAGE

☛ *Objectifs*

- ⇨ définir la notion de sondage et celle de recensement
- ⇨ présenter les différentes méthodes de sondage
- ⇨ introduire les méthodes aléatoires
- ⇨ définir l'échantillon et ses caractéristiques
- ⇨ donner les distributions des paramètres de l'échantillon.

☛ *Mots clés*

Sondage – recensement – méthodes des unités-types – méthodes des quotas – méthodes aléatoires – table des nombres au hasard – distribution d'échantillonnage : fréquence ; moyenne ; variance.

I ***Echantillonnage***
II ***Construction d'un échantillon***
III ***Distribution d'échantillonnage***
IV ***Distribution des paramètres***
V ***Intervalle de pari***

AUTO-EVALUATION & EXERCICES CORRIGES

L'analyse statistique d'un phénomène observé s'avère souvent difficile, soit pour des raisons de coût, voire même pour des raisons d'impossibilités matérielles.

Pour pallier ces difficultés et obtenir des informations à l'aide d'indicateurs fiables, on a recours à ***l'inférence statistique*** qui *« est un ensemble de méthodes basées sur le calcul des probabilités, qui permettent de juger des propriétés d'une population connue, inconnue ou incomplètement déterminée, au moyen d'un échantillon prélevé de cette population»*.

I - ECHANTILLONNAGE

I - 1 INTRODUCTION

Pour *clarifier* les idées et mieux ressortir les *concepts* sur lesquels repose l'inférence statistique, donnons deux exemples.

Exemple 1

Considérons les réponses X de N individus à une question qui admet deux réponses :

oui : on lui attribue par convention la valeur *1* ;

non : on lui attribue par convention la valeur *0* ;

Appelons p *la fréquence des « oui »* dans l'ensemble des réponses.

Deux cas de figures peuvent se présenter :

① effectuer *un recensement* (questionner la population entière) p est alors déterminé, mais à quel prix ?

② ou bien, il est *impossible de le faire* et la valeur de p est alors inconnue. Dans ce cas, pour avoir une idée de sa valeur, on procède comme suit :

- on extrait une partie **« un échantillon »** de l'ensemble des réponses de taille $\mathbf{n < N}$.
- à partir de cet échantillon, se posent plusieurs *types de problèmes du genre :*

◆ *Comment choisir l'échantillon ? Comment estimer la proportion* p *? Comment mesurer la précision de cette proportion ? Comment tester une hypothèse telle que la proportion de oui soit supérieure ou égale à une valeur donnée ?*

Exemple 2

La population est composée de lampes électriques fabriquées dans une grande usine. On se propose d'étudier *la durée de vie des lampes.*

La détermination de la distribution de cette population est impossible car elle nécessite leur destruction. On est alors amené à associer une **variable aléatoire continue** X « *durée de vie d'une lampe électrique.*» A cette variable, on assigne une distribution théorique caractérisée par une **fonction de fréquence** $F(x,\theta)$ où θ est un **paramètre** représentant une **valeur typique** de la population (moyenne, écart-type...)

Comme dans l'exemple précédent, à partir d'un échantillon extrait de la population, il s'agit d'apporter un **ensemble de réponses** :

- **estimer** le paramètre θ ;
- **vérifier** si les données de l'échantillon sont **compatibles** avec la distribution associée ;
- enfin **mesurer la précision** de l'estimation trouvée en lui associant un **intervalle**.

☞ Les deux exemples précédents montrent *qu'il s'agit d'utiliser des échantillons aux fins d'avoir des informations* sur le *phénomène observé* dont les *caractéristiques* sont largement *inconnues.*

☞ La *recherche de l'information* demeure un aspect essentiel de toute démarche. Outre qu'elle permet la *vérification des théories*, elle est devenue également un *outil de gestion* nécessaire à la prise de *décision*, qu'elle soit individuelle ou collective.

Ce chapitre traitera de certains aspects pratiques tels les **sondages** et les **distributions d'échantillonnage** dont le rôle est important dans **l'inférence statistique** qui fournit une réponse satisfaisante aux interrogations posées par les deux exemples précédents.

I - 2 SONDAGE ◆ RECENSEMENT

Considérons une population P, sur laquelle on veut rassembler un ensemble d'informations, (les réponses affirmatives du premier exemple ou la durée de vie du second exemple). Comme nous l'avons déjà signalé deux méthodes peuvent être employées :

* **Le recensement** : le mot lui-même évoque le décompte (*dénombrement*) du nombre des individus de la population soumis à enquête, d'où le caractère exhaustif de l'opération statistique. Le véritable recensement demeure celui de la population.

* **Le sondage** ou **l'échantillonnage** : Pour obtenir des *informations* sur la population, on en *prélève une partie* et on *l'examine* à fin d'obtenir les renseignements souhaités.

Chacune de ces deux méthodes présente des *avantages* et des *inconvénients*. Le *choix* dépend essentiellement du *phénomène étudié*. Dans ce qui suit, nous allons nous intéresser, *principalement aux sondages*, plusieurs raisons en justifient le recours.

I - 3 SONDAGE

La *méthode des sondages* a connu un *usage remarquable*, il suffit de penser au sondage d'opinion, aux enquêtes de marché, aux méthodes statistiques employées en contrôle de qualité... Cette méthode présente des *avantages* et des *inconvénients*.

I - 3. 1 Avantages

On peut signaler :

✓ **Le coût** : c'est un avantage déterminant. En effet, il est *moins coûteux* de faire une *étude sur une sous-population* que sur une population entière.

✓ **La rapidité** : l'objectif recherché est d'avoir une *information rapide* pour prendre des *décisions à court terme*. Les délais d'un *recensement* étant trop *longs*. Signalons à cet égard que les résultats du dernier recensement en Tunisie (*1994*) ne sont pas encore connus dans leur totalité. Dans des études conjoncturelles (opinion des chefs d'entreprise sur l'investissement par exemple) le recensement, même s'il est matériellement réalisable, serait inefficace.

✓ **L'impossibilité de recensement** comme dans le cas du deuxième exemple celui des lampes. L'examen de leur durée de vie aboutit à leur destruction !

II - 3. 2 Inconvénients

On peut citer à titre *non exhaustif* :

✓ **Les erreurs d'échantillonnage** :

* On peut introduire des *biais* dans l'estimation des caractères puisqu'on ne s'intéresse qu'à une partie de la population. Si l'échantillon retenu n'est pas représentatif de l'ensemble de la population par exemple, il peut exister un écart entre le résultat du sondage et celui qu'on aurait pu obtenir par un recensement.

* La méthode des sondages ne fournit pas des informations très détaillées.

✓ **Les erreurs d'observation :**

* Au niveau des erreurs de saisie de l'information : le recueil de l'information peut se faire par téléphone, entretien, questionnaire...voire par des mesures lorsqu'il s'agit d'unités physiques, ...

* La performance dépend de la conduite de l'enquête allant de la connaissance des objectifs jusqu'à l'exploitation des résultats.

* A chaque niveau du déroulement de l'enquête des erreurs et des interventions risquent de l'orienter et de la biaiser.

✓ **L'incertitude :**

* L'incertitude est liée à *l'interprétation* et à la *confiance* des résultats obtenus.

* Les variables étudiées dépendent de *comportements difficiles* à cerner et souvent *non permanents* : les opinions politiques, l'attitude des consommateurs, les pratiques d'entreprises...

✓ **La mauvaise interprétation des résultats :**

Cette erreur peut provenir de plusieurs sources :

* *L'échantillonnage lui-même :* On *ne peut affirmer* que l'échantillon retenu représente véritablement l'ensemble étudié (un député est-il représentatif de son électorat ?)

Si l'on a sondé *100 personnes* dans *l'exemple 1*, un résultat de *49%* de *oui, 51% de non n'est pas significatif.* La faible différence est inférieure aux fluctuations moyennes entre les différents sondages possibles qui est de l'ordre de 3%.

* *La taille de l'échantillon* : On peut considérer que l'erreur sera d'autant plus faible que l'échantillon sera plus grand. Dans le même exemple, pour atteindre une précision de 1%, il faut interroger un échantillon 9 fois plus grands.

* *La manière pratique dont l'échantillon a été choisi* :

- *Le tirage* a-t-il été réalisé de manière à représenter la réalité ?

- Comment réagir lorsque les *personnes ne répondent pas* à une question ?

- *La mesure des variables* elle-même, est une source d'erreurs, *citons à titre indicatif :*

- le choix des questions,
- la longueur du questionnaire,
- l'utilisation d'un langage trop technique ou inadapté,
- la mauvaise classification des questions,
- le dosage de la forme des questions entre les questions aux choix multiples et les réponses par oui ou par non,
- l'influence de l'enquêteur,
- le recours à des réponses suggérées,
- La tendance naturelle à dire *oui* indépendamment de la nature de la question (cette tendance est à l'origine de *10* à *12%* d'un supplément de *oui*).
- la précision de la mesure (âge en années, en mois...)

II - CONSTRUCTION D'UN ECHANTILLON

Considérons une population composée de N **individus** dont on étudie **un** ou **plusieurs caractères**. Il existe *plusieurs méthodes* pour choisir un **échantillon** de cette population. La méthode la plus courante est celle des sondages. On peut en distinguer *différents types* suivant que *l'échantillon* a été obtenu par **des méthodes empiriques** ou par **des méthodes aléatoires.**

II - 1 METHODES EMPIRIQUES

Ces méthodes reposent sur *l'expérience* et *le sérieux* de l'équipe qui réalise l'étude. Elles sont faciles à mettre en œuvre et sont applicables dans le cas de *très petits échantillons.*

II- 1.1 La méthode des unités-types

* Elle *dépend* principalement du *statisticien* et de sa connaissance approfondie de la population étudiée. Son principe consiste à prélever dans la population étudiée les divers individus de l'échantillon suivant des *critères fixés à priori.*

* La population est divisée en sous-ensembles homogènes et chacun d'eux représenté par quelques « **unités-types**».

* Ces unités-types doivent être réexaminées à chaque enquête.

Remarque

☞ Cette méthode est utilisée, généralement, quand les *moyens* matériels ne sont pas très importants.

II- 1. 2 La méthode des quotas

∗ Considérons *un caractère* ***X*** dont on veut étudier *la répartition* dans une population ***P.*** Des études préalables ont permis de lier ***X*** avec d'autres caractères ***Y, Z***..., dont on connaît la distribution dans la population totale ***P.***

La méthode des quotas consiste à *imposer à l'échantillon prélevé une structure analogue à celle de la population totale* ***P,*** relativement aux caractères contrôlés : ***Y, Z***...

Chacune de ces variables de contrôle ***Y, Z***...doit posséder les propriétés suivantes :

- Sa distribution dans la population ***P*** connue.
- Son observation ne comporte ni difficultés ni erreurs.
- Une relation étroite la lie à ***X.***

Pour fixer les idées, supposons que l'on étudie *un produit de grande consommation.* La consommation de ce produit dépend des facteurs socio-économiques de la population :

l'âge, le sexe, la catégorie socio-professionnelle *(C.S.P.)*

D'après le *recensement démographique*, la répartition de la population selon ces critères est donnée par le tableau suivant :

Sexe	%	**Age**	%	**C.S.P**	%
Masculin	36	18-25	20	**P**atrons	14
Féminin	64	25-50	40	**C**adres supérieurs	6
		50-60	15	**C**adres moyens	20
		60 et plus	25	**O**uvriers	45
				Inactifs	15
Σ	100		100		100

☞ Le principe est de *diversifier* pour réduire les fluctuations relatives et de se rapprocher de l'adage *« il ne faut pas mettre tous ses œufs dans le même panier ».*

☞ La méthode consiste à *imposer aux enquêteurs* le choix des individus qu'ils interrogent, dans les proportions « ou *quotas* » définies, ci-dessus, pour chaque caractère.

Remarques

☞ Cette méthode, très *utilisée* dans les *études de marché,* a des **avantages** pratiques évidents :

♣ Elle est d'un *prix de revient* peu élevé : l'enquêteur a toute la liberté d'interroger qui il veut et, par suite, concentrer ses interviews dans une région géographique restreinte, ce qui limite les frais.

♣ Elle n'exige *pas de base de sondage*. Une base de sondage est *une liste sans omission, ni répétition des individus* susceptibles d'être interrogés.

☞ *Mais elle possède également des* **inconvénients :**

♠ *risque de biais:*

* les enquêteurs ont tendance à interroger les personnes appartenant à leurs relations ;

* il peut se révéler dans l'exemple de l'étude du produit de grande consommation que le *degré d'urbanisation*, non introduit explicitement dans les quotas, soit *significatif*.

♠ difficulté de contrôle des enquêteurs.

♠ impossibilité d'estimer la précision des résultats.

II - 1.3 La méthode de Politz

C'est une *amélioration de la méthode des quotas* . Elle impose aux enquêteurs, pour chaque interview, un itinéraire.

Remarques

☞ Toutefois le plan de travail indiqué est rarement respecté. Pour remédier à cet inconvénient, quand les écarts ne sont pas trop élevés, on effectue des redressements.

☞ Ces redressements peuvent devenir compliqués lorsqu'il s'agit de plusieurs variables de contrôle.

II - 2 METHODES ALEATOIRES

A l'inverse des méthodes précédentes, les individus sont choisis par un *tirage au hasard* dans la population sondée. Dans ce cas, le **tirage est dit aléatoire.**

Chaque élément a *une probabilité* d'appartenir à l'échantillon. *L'équiprobabilité* des individus interrogés est assurée par ce que l'on a appelé une **base de sondage :** *c'est-à-dire une liste de toutes les unités qui constituent la population, sans omission, ni répétition.* Une telle liste est souvent *impossible à constituer* puisqu'elle est coûteuse à établir et à maintenir à jour. Pour *remédier à cette difficulté* et éviter de très gros risques d'erreurs, d'*autres méthodes plus au moins empiriques* permettent de déterminer des échantillons. Il s'agit dans ce qui suit de les examiner.

II - 2. 1 Sondage aléatoire simple

Il s'agit de constituer directement les échantillons à partir d'une base de sondage. **Deux conditions** doivent être satisfaites :
Les éléments sont :

❶ *extraits* de **manière aléatoire** c'est-à-dire chaque élément doit avoir la même probabilité d'appartenir à l'échantillon

❷ *tirés* **indépendamment les uns des autres** c'est-à-dire la probabilité d'extraction d'un élément ne doit pas dépendre des probabilités d'extraction des autres.

Lorsque ces deux conditions sont remplies, **l'échantillon** est dit **aléatoire simple.**

☞ Pour *construire un échantillon aléatoire simple à partir d'une base de sondage, deux variantes sont possibles.* Le tirage des unités peut être :

① **Sans remise** ou **tirage exhaustif** :

Les unités tirées successivement ou ensemble *ne sont pas remises* dans la population. Chaque unité figure *au plus une fois* dans l'échantillon.

② **avec remise** ou **tirage indépendant** :

Chaque unité tirée est observée puis *remise* dans la population avant qu'une autre ne soit tirée. Chaque unité peut *figurer plus d'une fois* dans l'échantillon.

Remarques

☞ Le tirage sans remise est le plus fréquent.

☞ Lorsque *la taille d'échantillon est petite* par rapport *à la taille de la population, les deux modes* de tirage sont *équivalents.*

II - 2. 2 Construction pratique d'un échantillon

Plusieurs pratiques sont utilisées pour construire un échantillon, citons les modalités les plus usuelles :

- la méthode du **tirage au sort** à l'aide de la **table des nombres au hasard**
- la méthode de **tirage systématique**.

a) **La table des nombres au hasard**

L*es nombres au hasard* désignent une série de chiffres obtenus indépendamment les uns des autres, chacun ayant une probabilité $\frac{1}{n}$ de sortie, *n* étant le *nombre de chiffres* du *système de numérotation* employé.

Cas particulier : *n = 10*

La table est constituée des nombres *0, 1, 2, 3, 4, 5, 6, 7, 8, 9 et chacun a la même probabilité d'apparition.* C'est un échantillon de la loi ***discrète uniforme*** (pour plus de détails sur la loi uniforme se reporter au manuel **Statistique & Probabilités**).

N'importe quel nombre de la table n'a aucune relation avec le nombre qui lui est au-dessus, en dessous, à sa droite ou à sa gauche.

Pour choisir des *nombres de la table*, il faut *procéder* comme suit :

① ***choisir un point d'entrée dans la table***

② ***choisir un itinéraire de lecture. On peut lire les nombres :***

- ***en ligne, de gauche à droite ou de droite à gauche***
- ***ou en colonne, de bas en haut ou de haut en bas.***
- ***on peut sauter un nombre sur deux***...

Exemple

Pour illustrer l'utilisation de la table, donnons un exemple :

N = 200 ***n = 8*** nature du *tirage : exhaustif*

On numérote chaque individu dans la base de sondage de *001* à *200* puis on choisit de la table des nombres de trois chiffres.

Pour *lire dans la table*, on utilise la *règle suivante* :

❶ Partant de la première ligne et ne considérant que les trois derniers chiffres de la quatrième colonne avec lecture de haut en bas.
❷ Ne retenir que les résultats compris entre *100* et *200* inclus
❸ Rejeter tout nombre apparaissant à nouveau dans la procédure de sélection
❹ Les individus portant les numéros :
109 045 075 157 199 026 051 118
constituent l'échantillon de ***taille n = 8***

b) **Le tirage systématique**

Lorsque la population est très grande, la *numérotation* devient très *laborieuse*. Il devient alors *difficile* de construire un échantillon au moyen d'une *table des nombres au hasard*. On ***prélève systématiquement*** les individus régulièrement espacés suivant un « ***pas*** » choisi. Si N est *la taille de la population*, n la *taille de l'échantillon*, on procède comme suit :

① On calcul le rapport $k = \frac{N}{n}$ c'est le *pas* à respecter entre les deux tirages.

② On choisit d'une *façon aléatoire* le *premier individu* dont le *numéro* se situe entre *1* et k

③ L'échantillon est ensuite constitué *en ajoutant successivement k.*
Si a est le premier numéro choisi, l'échantillon de taille n sera composé des individus du rang :

$$a,\ a + k,\ a + 2k,\ldots,\ a + (n\text{-}1)k$$

☞ Cette méthode, simple et rapide, est *particulièrement indiquée* dans le cas *de dossiers* ou *de listes*.

Exemple

Pour *illustrer* le tirage systématique, utilisons *l'exemple suivant* :

$$N = 200 \qquad n = 10 \qquad k = \frac{200}{10} = 20$$

A l'aide de *la table de nombres au hasard*, on choisit un nombre entier entre *01* et *20*

① Point d'entrée de la table : $11^{ème}$ ligne, $1^{ère}$ colonne
② Lecture de haut en bas de nombres de 2 chiffres (les deux premiers)
③ Nombre sorti : *14*

④ L'échantillon sera constitué par :

014 034 054 074 094 114 134 154 174 194

Remarque

☞ Le *coût élevé* d'un sondage, réalisé par ces méthodes, explique qu'on lui préfère des méthodes non aléatoires : la *méthode des quotas* par exemple.

II – 2. 3 Des méthodes améliorées

Des améliorations ont été apportées à la méthode aléatoire élémentaire, sans pour autant sacrifier le caractère probabiliste. Citons parmi elles :

- ♣ la stratification
- ♣ le sondage en grappes
- ♣ le sondage à plusieurs degrés
- ♣ le sondage avec des probabilités inégales

Nous allons, dans ce qui suit, les présenter, sommairement.

a) La stratification

Il s'agit de répartir les unités de la population en groupes homogènes, appelés **strates** suivant des caractères ***Y, Z***...supposés en corrélation avec ***X***. A l'intérieur de *chaque strate*, on prélève un *échantillon aléatoire*.

Remarques

☞ Cette méthode utilise les mêmes principes que la méthode des quotas mais en diffère par son caractère probabiliste.

☞ Elle permet d'accroître la précision pour une même taille de l'échantillon global.

b) Le sondage en grappes

Au lieu de prélever les unités statistiques isolément, on *prélève plusieurs simultanément*. On décidera, par exemple, d'interroger tous les ménages d'un même ***immeuble*** ce qui constitue une grappe, les ***ménages*** étant les *grains*. Le procédé consiste à prélever au hasard un certain nombres de grappes et à choisir toutes les unités de chaque grappe.

*c) **Le sondage à plusieurs degrés***

Il consiste à réaliser des tirages en « *cascades* ». Pour mieux le comprendre, reprenons le ***sondage à deux degrés*** utilisé par **l'Institut national de statistique** (*I. N. S*) dans *l'enquête nationale* sur *le budget et la consommation des ménages (juin 1990 - juin 1991)* :

« La base du sondage est constituée par un échantillon de 1120 zones géographiques tirées du fichier du recensement de la population mise à jour en 1989. Cette base du sondage est stratifiée selon deux critères le gouvernorat et le milieu d'habitat...

✧ *Au premier degré*, un échantillon de 666 unités primaires avec une *probabilité proportionnelle à la taille des districts* (en nombres de ménages)

✧ *Au deuxième degré* et dans chaque *district-échantillon*, un tirage de *12* ménages échantillons est effectué avec une probabilité égale en tenant compte de la taille du ménage et du nombre d'actifs occupés par ménage.

✧ La taille de l'échantillon tiré est de *8000*. Mais le nombre de ménages effectivement enquêtés s'élève à *7734* soit *96,8%* de l'échantillon initial du tirage. Cet écart est la conséquence des cas de non-réponses soit pour changement de domicile ou de refus de réponses ou pour circonstances exceptionnelles (problèmes familiaux, décès...) ».

*d) **Le sondage avec des probabilités inégales***

Au lieu de tirer les unités du sondage en donnant à chacune la même probabilité d'appartenir à l'échantillon, on peut leur affecter des *probabilités différentes.* C'est le *cas* décrit précédemment celui *de l'enquête nationale sur le budget et la consommation des ménages.*

Remarques

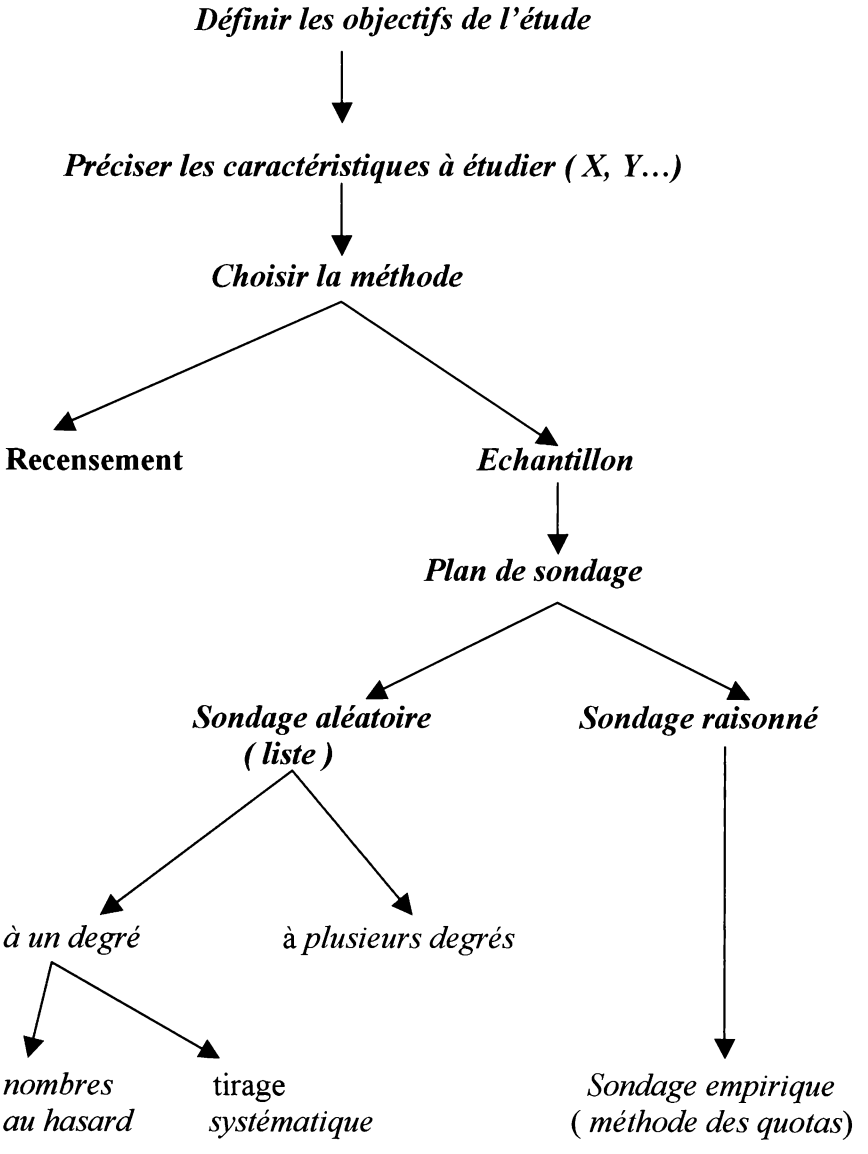
☞ Le coût sera plus faible que pour un sondage à un seul degré.

☞ A taille égale, la précision sera moindre.

☞ Les unités primaires doivent être assez nombreuses (districts).

☞ Elles doivent être d'importance comparable. Sinon les unités primaires seront tirées avec des probabilités inégales et proportionnelles à leur taille.

La pratique des sondages



III - DISTRIBUTION D'ECHANTILLONNAGE

Pour mieux comprendre les distributions d'échantillonnage, il est commode de faire appel à *un exemple simple* puis de le *généraliser*.

III - 1 EXEMPLE

On dispose d'une **population** ***P*** de taille ***N = 4*** dont les valeurs sont *2,4,6,8*

▪ La ***moyenne*** de cette population est donnée par $m = \frac{\sum_{1}^{4} x_i}{4} = 5$

▪ **L'écart-type** par : $\sigma = \sqrt{\frac{\sum_{1}^{4} (x_i - 5)^2}{4}} = 2{,}23$

On tire un **échantillon** de cette population de taille ***n = 2***

Ce tirage peut être **avec remise** (l'élément prélevé est remis dans la population après avoir noté sa valeur). A chacun des quatre éléments, on peut associer l'un des quatre autres, on obtient alors *16 cas possibles donnés* par le tableau suivant :

	2	4	6	8
2	(2,2)	(2,4)	(2,6)	(2,8)
4	(4,2)	(4,4)	(4,6)	(4,8)
6	(6,2)	(6,4)	(6,6)	(6,8)
8	(8,2)	(8,4)	(8,6)	(8,8)

♦ Calculons la moyenne de chaque échantillon x_i, on obtient respectivement le tableau suivant :

2	3	4	5
3	4	5	6
4	5	6	7
5	6	7	8

• Ce sont des moyennes différentes de celle de la population

• La *différence* par rapport à *la moyenne* de la population est *plus ou moins grande.*

• La *distribution des valeurs* que peut prendre la *moyenne d'échantillon* sur tous les échantillons possibles, est appelée ***distribution d'échantillonnage.***

• Calculons la *moyenne des moyennes*, on obtient : $\bar{x} = m = 5$

$\Rightarrow$ ***La moyenne des moyennes est égale à la moyenne de la population***

♦ Calculons ***l'écart-type de cet ensemble de moyennes*** par rapport à ***la moyenne de la population :***

$$\sigma_x = \sqrt{\frac{\sum_{1}^{16}(x_i-5)^2}{16}} = \sqrt{2,5} = 1,58$$

$$\Rightarrow \sigma_x = \frac{\sqrt{5}}{\sqrt{2}} = \sqrt{2,5} \text{ d'où } \sigma_x = \frac{\sigma}{\sqrt{n}}$$

♦ Dans le cas d'un tirage ***sans remise*** (l'élément prélevé, après avoir noté sa valeur, est définitivement retiré de la population) **l'écart-type** devient :

$$\sigma_x = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \; ; \quad \sqrt{\frac{N-n}{N-1}} \text{ est appelé } \textit{facteur d'exhaustivité}$$

Si on utilise une grande population, N tend vers l'infini, le facteur d'exhaustivité tend vers 1.

III - 2 GENERALISATION

L'expérience aléatoire consiste généralement dans le *prélèvement au hasard* d'un individu dans une ***population P*** « *population mère* ou *parente* » et d'*observer* pour chaque individu *la valeur de la caractéristique*. Les résultats de *n* répétitions de cette expérience peuvent être interprétés comme un *échantillon aléatoire*.

Chaque échantillon constitue une *distribution statistique* pour laquelle on peut calculer *une ou plusieurs valeurs typiques* : *moyenne, mode, écart-type, proportion*...

Cette valeur typique calculée va varier *d'un échantillon à un autre* (cf. *l'exemple*).

Pour la *totalité des échantillons prélevés* dans la population ***P,*** on va obtenir *plusieurs valeurs* de cette *valeur typique*. On est alors en présence d'une *distribution d'échantillonnage* : distribution d'échantillonnage d'une *proportion*, distribution d'échantillonnage d'une *moyenne*, distribution d'échantillonnage d'une *variance*...

Il est possible *de résumer la distribution d'échantillonnage* par des valeurs *typiques* appelées *valeurs typiques de l'échantillon* ou *valeurs types des valeurs observées*.

Connaissant la distribution de ***X****, peut-on en déduire des informations sur les caractéristiques de l'échantillon ?*

Existe-t-il des relations entre les caractéristiques des échantillons et celles des distributions d'échantillonnage ?

Remarques

☞ Rappelons que nous n'avons pas fait de distinction entre les deux termes *loi de probabilité* d'une variable aléatoire (*calcul des probabilités*) et *distribution* d'une variable (*statistique mathématique*) puisqu'ils sont *synonymes. Il en est de même*, nous avons utilisé *densité de probabilité* d'une variable connue ou *fonction de fréquence* de la variable (statistique mathématique).

☞ Les résultats de *n répétitions* d'une *expérience aléatoire* peuvent être interprétés comme un *échantillon aléatoire* prélevé d'une *population infinie* de toutes *les observations possibles.*

☞ *Les valeurs observées* $x_1,...,x_n$ peuvent être considérées comme un ensemble de *valeurs-échantillon indépendantes* de la variable aléatoire ***X***. La *distribution* de ***X*** est la *distribution de la population* ou *la distribution théorique du phénomène* considéré. Elle peut être *définie* par la donnée de la *fonction de répartition*, de la *fonction densité de probabilité*...

☞ Aux valeurs observées $x_1,...,x_n$, on associe à chacune d'elle une probabilité de $\frac{1}{n}$.

L'ensemble $\{(x_i, \frac{1}{n})\}$ est une distribution *de données statistiques* ou *une distribution de l'échantillon* ou *une distribution théorique.*

Considérons, à nouveau, le cas d'une population ***P*** de moyenne ***m*** et d'écart-type σ, on prélève de cette population un ***échantillon de taille n*** :

✧ L'échantillon prélevé doit être :

➢ *représentatif* c'est-à-dire qu'il possède une certaine *ressemblance* avec la population dont il est extrait,

➢ *aléatoire* c'est-à-dire que ses *éléments sont pris au hasard.*

✧ L'échantillon prélevé peut être également :

➢ *avec remise* ou *non exhaustif* c'est-à-dire l'élément *prélevé* au premier tirage est *remis* dans la population mère.

➢ *sans remise* ou *exhaustif* c'est-à-dire l'élément *prélevé* est définitivement *retiré* de la population mère.

✧ *La distribution d'échantillonnage* caractérise les *fluctuations* d'échantillonnage de tout *paramètre* calculé sur les échantillons possibles de *même taille.*

✧ *Le paramètre mesuré* sur chaque échantillon de taille ***n*** sera une variable aléatoire qui

* prend *diverses valeurs* selon les résultats de l'échantillonnage

* possède *une distribution de probabilité* (*densité de probabilité, fonction de répartition..*)

Notation

On désignera par :

N la *taille de la population* mère (***N*** peut être égal à l'infini)

m la *moyenne* de la variable dans la *population* mère. Elle est généralement inconnue et l'on cherche à l'estimer ;

σ *l'écart-type* (la dispersion des individus autour de ***m***) ;

n la *taille de l'échantillon* prélevé dans cette population ;

v_i le *nombre d'individus* ayant la valeur x_i du caractère ;

f_i la *fréquence* d'une valeur x_i dans l'échantillon $\quad f_i = \frac{v_i}{n}$

$\overline{X}$ la *moyenne* de la variable dans *l'échantillon* $\quad \overline{X} = \frac{1}{n}\sum_1^n X_i$

s l'écart-type de l'*échantillon* (la dispersion autour de ***m***) :

$$s = \sqrt{\frac{1}{n}\sum_1^n (X_i - \overline{X})^2}$$

S^2 la *quasi-variance* $\quad S^2 = \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2$

La distribution d'échantillonnage caractérise les fluctuations des paramètres remarquables : fréquence, moyenne, variance... Après avoir rappelé des *théorèmes importants de l'espérance et de la variance*, on procédera à *l'étude des distributions de ces paramètres.*

IV - DISTRIBUTION DES PARAMETRES

Dans l'étude des distributions d'échantillonnage, les deux théorèmes suivants sont très utiles :

IV - 1 THEOREMES

❶ Théorème I

L'espérance mathématique d'une somme de variables aléatoires est égale à la somme des espérances mathématiques :

$Y = X_1 + ... + X_i + ... + X_n$ *alors* $E(Y) = E(X_1) + ... + E(X_i) + ... + E(X_n)$

❷ Théorème II

La variance d'une somme de variables aléatoires indépendantes est égale à la somme des variances :

$$Y = X_1 + ... + X_i + ... + X_n \text{ où } X_1, ..., X_i ..., X_n$$

sont indépendantes alors :

$$V(Y) = V(X_1) + ... + V(X_i) + ... + V(X_n)$$

☞ On utilise également les résultats ci-dessous démontrés dans le manuel « ***Statistique & Probabilités*** » et également rappelés au ***chapitre I*** de cet ouvrage.

$$E(aX + b) = aE(X) + b \text{ et } V(aX+b) = a^2 V(X)$$

IV - 2 DISTRIBUTION D'UNE FREQUENCE

A partir de la population ***P***, on prélève un échantillon ainsi déterminé :

* le premier ***tirage*** étant une ***expérience aléatoire***, on lui associe une variable aléatoire X_1 qui peut prendre les valeurs $x_1, ..., x_i, ..., x_n$ avec les probabilités $p_1, ..., p_i, ... p_n$.

* il en de même du $i^{ème}$ et du $n^{ième}$ tirage.

Les ***n*** tirages successifs ainsi constitués $(X_1, ..., X_i ..., X_n)$ forment l'échantillon de taille ***n***. Les tirages étant non exhaustifs, les ***n*** variables aléatoires sont *indépendantes.*

Soit v_i le nombre d'individus ayant la valeur x_i du caractère dans l'échantillon prélevé dans la population ***P***, le tirage étant *avec remise*, la distribution de v_i est *une distribution binomiale* de paramètres *m* et σ.

La fréquence f_i d'une valeur x_i dans l'échantillon est définie par $\frac{v_i}{n}$.

Il s'agit de déterminer *l'espérance* et *la variance* de cette fréquence :

❶ Espérance

$$E(f_i) = E(\frac{v_i}{n}) = \frac{1}{n} E(v_i) = \frac{1}{n} n p_i = p_i \quad \Rightarrow \quad E(f_i) = p_i$$

❷ Variance

$$V(fi) = V(\frac{v_i}{n}) = \frac{1}{n^2} npi(1-p_i) = \frac{p_i(1-p_i)}{n} \quad \Rightarrow \quad V(fi) = \frac{p_i(1-p_i)}{n}$$

Remarques

☞ La moyenne de la fréquence f_i dans l'échantillonnage est égale à p_i

☞ La variance de f_i tend vers zéro lorsque la taille ***n*** de l'échantillon devient infinie.

☞ Pour ***n*** *grand*, la distribution de l'échantillon est voisine de celle de la *population totale.*

☞ Si ***n*** *est grand* la distribution binomiale de f_i tend vers *une distribution normale*.

IV - 3 DISTRIBUTION DE LA MOYENNE

Il s'agit de déterminer *l'espérance de la moyenne* et *sa variance*.

La moyenne de l'échantillon est définie par la relation :

$$\overline{X} = \frac{X_1 + ... + X_i + ... + X_n}{n}$$

Par application du *théorème I,* on a : $E(\overline{X}) = \frac{1}{n} E(\sum_1^n X_i) = \frac{1}{n} \sum_1^n E(X_i)$

Or $E(X_1) = ... = E(X_i) ... = E(X_n) = m$ d'où $E(\overline{X}) = m$

Par application du *théorème II,* on a :

$$V(\overline{X}) = \frac{1}{n^2} V(\sum_1^n X_i) = \frac{1}{n^2} \sum_1^n V(X_i)$$

Or $V(X_1) = ... = V(X_i) = ... = V(X_n) = \sigma^2$ d'où $V(\overline{X}) = \frac{\sigma^2}{n}$

Remarques

☞ Lorsque l'échantillonnage s'effectue sans remise, on doit apporter une correction à la variance. Dans ce cas : $V(\overline{X}) = \frac{\sigma^2}{n}(\frac{N-n}{N-1}) \# \frac{\sigma^2}{n}(1-\frac{n}{N})$

☞ $\frac{n}{N}$ représente le taux de sondage, lorsqu'il est *inférieur ou égal à 0,05*, on peut l'ignorer.

VI - 4 DISTRIBUTION DE LA VARIANCE

On sait que la ***variance de l'échantillon*** s^2 est une *variable aléatoire*, reliée aux variables aléatoires $X_{1,\ldots,} X_{n,}$ par la relation :

$$s^2 = \frac{1}{n}\sum_{1}^{n}(X_i - \overline{X})^2$$

Elle est appelée également ***écart quadratique moyen.*** Il s'agit de déterminer son *espérance* et sa *variance*.

❶ Espérance $E(s^2) = E\left[\frac{1}{n}\sum_{1}^{n}(X_i - \overline{X})^2\right]$

Or d'après le *théorème de Kœnig*, on démontre que ***l'espérance est donnée par*** $E(s^2) = \frac{n-1}{n}\sigma^2$

❷ Variance

La variance a une expression complexe. On démontre que :

$$V(s^2) = \frac{\mu_4 - \mu^2_2}{n} - 2\frac{\mu_4 - 2\mu^2_2}{n^2} + \frac{\mu_4 - 3\mu^2_2}{n^3}$$

$$V(s^2) \to 0; \quad \text{lorsque } n \to \infty$$

Remarque

☞ En général, pour définir la variance de l'échantillon, on utilise l'expression $S^2 = \frac{1}{n-1}\sum_{1}^{n}(X_i - \overline{X})^2$. Elle se prête mieux aux calculs et a des propriétés remarquables. Ses *caractéristiques* sont :

$$E(S^2) = \sigma^2$$

$$V(S^2) = \frac{\mu_4 - 3\mu^2_2}{n^3} + \frac{2\mu_2^2}{n-1} = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4)$$

$$V(S^2) \to 0; \quad \text{lorsque } n \to \infty$$

V - INTERVALLE DE PARI

V - 1 DEFINITION

C'est un *intervalle* IP_α dans lequel on doit s'attendre à trouver la valeur de la variable observée, avec un *risque* α de se *tromper*.

Remarque

☞ Cet intervalle peut être centré sur la valeur moyenne m de la variable X, telle que la probabilité pour que la variable X appartienne à cet intervalle soit égale à $1-\alpha$: $\text{Prob}(X \in IP_\alpha) = \text{Prob}(m - a \leq X \leq m + a) = 1 - \alpha$

V - 2 CONSTRUCTION

V - 2. 1 Pour une loi normale: On sait que la *variable centrée réduite* $T = \frac{X-\mu}{\sigma} \rightarrow N(0,1)$ d'où $\text{Prob}(-\frac{a}{\sigma} \leq T \leq \frac{a}{\sigma}) = 1 - \alpha$.

En posant $t_\alpha = \frac{a}{\sigma}$, on a : $\text{Prob}(X \in IP) = \text{Prob}(|T| \leq t_\alpha) = 1 - \alpha$

V - 2. 2 Pour une moyenne

Si $n \geq 30$ ou $X \rightarrow N(\mu, \sigma)$ alors : $\overline{X} \rightarrow N(\mu, \frac{\sigma}{\sqrt{n}})$; $IP_\alpha = [\, \mu \pm t_\alpha \frac{\sigma}{\sqrt{n}} \,]$

V - 2. 3 Pour une proportion

L'intervalle de pari est défini par *un écart* ***e*** autour de ***P***. Soit p_0 la variable d'échantillonnage, si les conditions de *normalité* sont satisfaites :

np et ***nq*** ≥ 15 alors $IP_\alpha = [\, p \pm t_\alpha \sqrt{\frac{pq}{n}} \,]$ pour p_0

Exemple

On parie que la proportion de boules blanches, sur un échantillon de 100 boules extrait d'une urne où $p = 50\%$, sera comprise entre 40% et 60%. Quelle est la probabilité de se tromper ?

L'écart autour de 50% est $e = 0,1$ d'où $\varepsilon = \frac{e}{\sigma} = \frac{0,1}{0,05} = 2$ ce qui correspond à $\alpha = 0,05$ (*lecture dans la table N(0, 1)*).

Il y a 5 chances sur dix de se tromper.

- Lorsque n augmente *l'intervalle de pari se rétrécit* autour de la valeur centrale.
- L'expression : $\varepsilon = \frac{e}{\sigma}$ avec $\sigma = \sqrt{\frac{pq}{n}}$ est appelée *écart-réduit*.

Auto-évaluation ?

1- Une liste qui contient sans omission, ni répétition des individus qui constituent la population s'appelle :

a) échantillon b) base de sondage.

2- Les méthodes pratiques pour constituer un échantillon sont :

a) le tirage au sort b) le tirage systématique c) un choix des unités.

3- La table des nombres au hasard est utilisée pour

a) le tirage au sort b) le tirage systématique

4- Pour une distribution d'échantillonnage prélevée dans une population infinie ayant une moyenne m et un écart-type σ, les paramètres sont :

a) $E(\overline{X}) = nm$ b) $E(\overline{X}) = m$

c) $Var(\overline{X}) = \sigma^2$ d) $Var(\overline{X}) = \frac{\sigma^2}{n}$

5- Un tirage est dit indépendant s'il est :

a) sans remise b) avec remise

6- Un tirage est dit exhaustif

a) sans remise b) avec remise

7- Lorsque le tirage est exhaustif, la variance de $\overline{X}$ est égale à :

a) $Var(\overline{X}) = \frac{\sigma^2}{n}$ b) $Var(\overline{X}) = \frac{\sigma^2}{n}(\frac{N-n}{N-1})$

8- La moyenne arithmétique observée sur chaque échantillon de taille n est :

a) une variable quelconque b) une variable aléatoire.

9- La table des nombres au hasard est caractérisée par le choix :

a) d'un point d'entrée dans la table

b) d'un intermédiaire de lecture

c) d'une base de sondage.

10- Lorsque la population est très grande, on utilise :

a) le tirage au sort b) le tirage systématique.

11- Un tirage systématique est caractérisé par le rapport :

a) $\frac{n}{N}$ b) $\frac{N}{n}$

Exercices ?

Exercice 1

On considère que le poids à la naissance est une variable aléatoire de moyenne m = 3300 g et d'écart- type σ = 500 g. On a prélevé un échantillon aléatoire de taille 50 parmi 100 dossiers de la maternité d'un hôpital A de Tunis.

Calculer les paramètres de la variable d'échantillonnage poids de naissance moyen.

Exercice 2

Calculer l'espérance de la variance de la variable $X \rightarrow N$ (m, σ) dans un échantillon de taille n ?

Exercice 3

Désignons par ***P*** la population tunisienne et notons par ***p*** la proportion d'individus de ***P*** possédant une automobile. Désignons par $\sum_n$ l'ensemble des échantillons de taille n, issu de ***P*** que l'on peut obtenir par sondages aléatoires (nous supposons que les n individus sont tirés au hasard l'un après l'autre, chaque individu étant remis dans la population avant le tirage de l'individu suivant).

Notons F_n la variable aléatoire définie par l'ensemble $\sum_n$: A tout échantillon de taille n , on fait correspondre la proportion f_n d'individus qui possèdent une automobile .

Pour tout $i \in \{1,2,..,n\}$, soit X_i la variable aléatoire définie sur l'ensemble $\sum_n$ telle que, étant donné un échantillon $E_n \in F_n$:

$$\begin{cases} X_i = 1 & \text{si le } i^{ème} \text{ individu de } E_n \text{ possède une voiture} \\ X_i = 0 & \text{sinon} \end{cases}$$

1-

a) Exprimer la variable aléatoire F_n en fonction des variables aléatoires $X_1, \ldots, X_n$

b) Quelle est la loi de probabilité de la variable aléatoire nF_n

c) Calculer $E(nF_n)$ et $Var(nF_n)$

d) Dans le cas où $n = 64$ et où l'on sait que p est tel que $0,4 \leq p \leq 0,6$, par quelle loi de probabilité peut-on approximer la distribution de la variable aléatoire nF_n ?

2-

a) Dans ce dernier cas (d), déterminer $t > 0$ tel que :

$$\text{Prob}\left(-t \leq \frac{F_n - E(F_n)}{\sqrt{Var\,(F_n)}} \leq t\right) = 0,95$$

b) Soit E_n un échantillon de taille $n=64$ sur lequel on a mesuré $F_n = 0,45$, déterminer alors un intervalle p_1 *et* p_2 tel que : $\text{Prob}\,(p_1 \leq p \leq p_2) = 0,95$

Concours d'accès au 2ème cycle ISG, Tunis 1990

Exercice 4

Soit $(X_1,\ldots,X_n)$ un échantillon de taille n indépendant et de même loi N (m, σ^2), on pose :

$$S^2 = \frac{\sum_{i=1}^{n}(X_I - \overline{X})^2}{n-1}$$

1- Calculer $E(S^2)$ et $Var(S^2)$

2- Si $n = 65$ et $S^2 = 16$, déterminer k tel que $P\,(S^2 \leq k) = 0,95$.

3- Si m = 20, déterminer t tel que $\text{Prob}\left(\frac{\overline{x} - 20}{S} > t\right) = 0,15$

Exercice 5

Soit $(X_1,\ldots,X_n)$ un échantillon de taille n pris dans une population dont le caractère X suit une loi de *Bernoulli* de paramètre p.

1- Calculer $E(\overline{X})$ et $Var(\overline{X})$

2- Montrer que $E\left[\frac{\overline{X}(1-\overline{X})}{n}\right] = \frac{n-1}{n^2}p(1-p)$

Exercice 6

Dans une population ***P*** la taille des individus est distribuée normalement autour d'une valeur moyenne $m = 1,70$ avec une dispersion $\sigma = 0,1$. On prélève au hasard 10 individus pour étudier cette répartition des tailles.

1- Quelles sont les expressions de la moyenne et de la dispersion observées : m'_1 et σ'_1

2- Si l'on prélève successivement un grand nombre d'échantillons identiques au précèdent, quelle est la moyenne des moyennes observées et la moyenne de leur dispersion.

3- Quelle est la moyenne des carrés des dispersions observées

4- Quelle est la dispersion des carrés des dispersions observées.

Exercice 7

Pour exploiter un fichier de 1000 malades, on décide de vérifier si les dossiers ont été correctement saisis, à l'aide d'un échantillon de 200 dossiers.

1- Comment doit-on procéder ?

2- On suppose que l'erreur commise, dans la saisie de l'ensemble des fichiers, est de 10%. Quelle est l'erreur commise dans l'échantillon ?

3- On a observé 25 dossiers mal saisis dans l'échantillon tiré. Que peut-on conclure sur l'hypothèse avancée ?

4- D'après les observations faites sur l'échantillon, à quel nombre de fichiers mal saisis doit-on s'attendre ?

Corrigés

Auto-évaluation

1- base de sondage

2- le tirage au sort ; le tirage systématique

3- le tirage au sort

4 - $E(\overline{X}) = m$; $Var(\overline{X}) = \frac{\sigma^2}{n}$

5- avec remise

6- sans remise

7- $Var(\overline{X}) = \frac{\sigma^2}{n}(\frac{N-n}{N-1})$ et $(\frac{N-n}{N-1})$ est un facteur de correction.

$Var(\overline{X}) = \frac{\sigma^2}{n}(1-\frac{n}{N})$; $\frac{n}{N}$ représente le taux de sondage et peut être négligé s'il est inférieur à 5% (voire même à 10 %)

8- une variable aléatoire

9- d'un point d'entrée dans la table ; d'un intermédiaire de lecture ; d'une base de sondage

10- le tirage au sort ; le tirage systématique

11- $\frac{N}{n}$ c'est l'intervalle fixe à respecter entre deux tirages. Il est appelé « *pas* ».

Exercices

Exercice 1

Calcul des paramètres

$$E(\overline{X}) = m = 3300$$

$$Var(\overline{X}) = \frac{\sigma^2}{n} = \frac{500^2}{100} = 2500$$

Exercice 2

Par définition : $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$ il s'agit de déterminer $E(s^2)$

On pose : $x_i - \overline{x} = (x_i - m) + (m - \overline{x})$

$$(x_i - \overline{x})^2 = (x_i - m)^2 + (m - \overline{x})^2 + 2(x_i - m)(m - \overline{x})$$

$$E(s^2) = E\left(\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}\right) = \frac{1}{n}E\left[(x_i - m)^2 + (m - \overline{x})^2 + 2(x_i - m)(m - \overline{x})\right]$$

En appliquant les propriétés de l'espérance, on obtient :

$$E(s^2) = \frac{n-1}{n}\sigma^2$$

Exercice 3

① a) *Détermination de la variable* F_n :

$$F_n = \frac{\sum_{1}^{n} X_i}{n}$$

b) $nF_n \in \{0,1,2,\ldots,n\}$ et nF_n suit une loi binomiale $B(n, p)$

nF_n est une somme de n variables de *Bernoulli,* toutes de même loi $B(1, p)$ et indépendantes (stabilité de la loi binomiale) ;

c) *Calcul des paramètres de la loi*

$$E(nF_n) = np \text{ et } \mathrm{Var}(nF_n) = np(1-p)$$

d) *Approximation de la loi* :

Les conditions de normalité étant vérifiées $n \geq 30$, $np \geq 15$ et $npq > 5$, on peut approximer la distribution par une *loi normale* $N(m, \sigma)$

② a) On sait que

$$T = \frac{nF_n - E(nF_n)}{\sqrt{Var(nF_n)}} \rightarrow N(0, 1)$$

$$\text{Prob}\left(-t \leq \frac{F_n - E(F_n)}{\sqrt{Var(F_n)}} \leq t\right) = 0,95 \Rightarrow t = 1,9$$

(par lecture sur la table de la loi normale centrée réduite)

b) *Détermination* de p_1 et p_2

$$\text{Prob}\left(-t \leq \frac{0,45-p}{\sqrt{\frac{p(1-p)}{64}}} \leq t\right) = 0,95$$

$$n = 64\,, \quad E(F_n) = p \text{ et } Var(F_n) = \frac{p(1-p)}{n}$$

Soit : $(0,45 - p)^2 \leq 0,06p(1-p) \Leftrightarrow 1,06p^2 - 0,96p + 0,2025 \leq 0$

En résolvant l'équation, on obtient :

$$p_1 = 0,33 \quad \text{et} \quad p_2 = 0,57$$

Exercice 4

$(X_1,\ldots,X_n)$ un échantillon de taille n indépendant et de même loi $N(m, \sigma^2)$

1- *Détermination des caractéristiques*

① *Calcul de* $E(S^2)$

$$E(S^2) = E\left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]$$

$$= \frac{\sigma^2}{(n-1)}E\left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}\right] = \frac{\sigma^2}{(n-1)}E(\chi^2_{n-1}) = \sigma^2$$

② *Calcul de Var* (S^2)

$$Var(S^2) = Var\left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}\right] = \frac{1}{(n-1)^2}Var\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]$$

$$\frac{\sigma^4}{(n-1)^2}Var\left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}\right] = \frac{\sigma^4}{(n-1)^2}Var(\chi^2_{n-1}) \quad \text{Soit :} \quad Var(S^2) = \frac{2\sigma^4}{n-1}$$

2- *Détermination de k*

1) n = 65 et $S^2 = 16$ $\qquad P(S^2 \le k) = 0,95$

$$\text{Prob}\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \le \frac{k(n-1)}{\sigma^2}\right] = 0,95$$

$$\text{Or} \quad \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \rightarrow \chi^2_{n-1}$$

On sait que lorsque le nombre de degré de liberté devient grand $v > 30$, la loi du khi-deux tend à se comporter comme une loi de Gauss et en particulier

$$\sqrt{2\chi^2} - \sqrt{2v-1} \rightarrow N(0,1)$$

d'où $\qquad \text{Prob}(\sqrt{2\chi^2} - \sqrt{127} < \sqrt{2A} - \sqrt{127}) = 0,95$

soit $\qquad \sqrt{2A} - \sqrt{127} = 1,65 \Rightarrow A \# 83,45 \qquad$ d'où $k \# 20,86$

Si $m = 20 \qquad \text{Prob}\left(\frac{\bar{x}-20}{S} > t\right) = 0,15$

On sait que : $\frac{\bar{x}-\mu}{\frac{S}{\sqrt{N}}}$ suit *la loi de Student* de degré de liberté *(n-1)*

Lorsque le nombre de degré de liberté devient grand ($n > 30$), la loi de Student tend à se comporter comme une loi de Gauss :

$$\text{Prob}(T > t\sqrt{65} = 0,15 \Rightarrow t\sqrt{65}) = 1,03 \qquad \text{soit} \quad t \# 0,128$$

Exercice 5

① Calcul de $E(\bar{X})$ et $Var(\bar{X})$

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = p \qquad Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{p(1-p)}{n}$$

② *Formule*

$$E\left[\frac{\bar{X}(1-\bar{X})}{n}\right] = \frac{1}{n}E\left[\bar{X}-\bar{X}^2\right] = \frac{1}{n}\left[E(\bar{X}) - Var(\bar{X}) - \left[E(\bar{X})\right]^2\right]$$

$$E\left[\frac{\bar{X}(1-\bar{X})}{n}\right] = \frac{1}{n^2}\left[np(1-p) - p(1-p)\right] \Rightarrow E\left[\frac{\bar{X}(1-\bar{X})}{n}\right] = \frac{n-1}{n^2}p(1-p)$$

Exercice 6

① *Expression de la moyenne et de la dispersion*

moyenne : $m'_1 = \frac{x'_1+x'_2+\ldots x'_{10}}{10}$

dispersion : $\sigma_1 = \sqrt{\frac{(x'_1-m'_1)^2+\ldots+\ldots(x'_{10}-m'_{10})^2}{10}}$

② *Moyenne des moyennes*

La moyenne des moyennes observées est égale à la moyenne de la population soit : 1,7

La moyenne de leur dispersion est égale à :

$$\frac{\sigma}{\sqrt{n}} = \frac{0,1}{3,16} \# 0,032$$

③ *Moyenne des carrés des dispersions* :

$$\frac{n-1}{n}\sigma^2 = \frac{9}{10} \times 0,01 = 0,009$$

④ *Dispersion des carrés des dispersions observées*

$$\frac{\sqrt{2(n-1)}}{n}\sigma^2 = 3\sqrt{2} \times 0,001 \# 0,0042$$

Exercice 7

① *Comment constituer l'échantillon ?*

L'échantillon de 200 dossiers doit être représentatif de l'ensemble des dossiers :

Population : le fichier des malades

Base de sondage : la liste des noms des individus dans le fichier

Unité statistique : les individus

Taille de l'échantillon : 200

Mode du tirage : sans remise

On commence par numéroter chaque individu dans la base de sondage de 0001 à 1000. On utilise la *table des nombres au hasard*. Pour lire cette table, on pique un nombre par exemple 37996 ($5^{\text{ième}}$ ligne, $2^{\text{ième}}$ colonne). Comme la base de sondage comporte 1000 individus, on ne retient que 4 chiffres on lit sur la ligne, on obtient 3799 ; 7846 ; 5720…

Les 200 premiers nombres ainsi obtenus sont ceux des dossiers qui constituent l'échantillon.

② *Il s'agit d'un problème de pari.*

On choisit un risque de 5% (généralement utilisé) $\alpha = 0,05$. Les conditions de validité de la loi normale *np* et *nq supérieurs à* 5 sont bien vérifiées. En effet : *np = 20 et nq =180*

$$\sigma = \sqrt{\frac{pq}{n}} = 0,0212$$

Une lecture de la table de la loi normale centrée réduite nous donne pour $\alpha = 0,05$, $t_\alpha = 1,96$ et l'intervalle de pari $IP_\alpha = 0,1 \pm 0,04$

Soit de 6 à 14%

③ La *fréquence observée* $\frac{25}{200} = 0,125$ appartient à l'intervalle précèdent, on ne peut rejeter l'hypothèse de 10%

On calcule :

$$t_\alpha = \frac{0,125 - 0,1}{0,0212} = 1,1792$$

Cette valeur est inférieure à 1,96 on ne peut rejeter l'hypothèse.

④ *Il s'agit de calculer*

$$p = p_0 \pm t_\alpha \sqrt{\frac{p_0\, q_0}{n}} = 0,125 \pm 0,0458$$

Il faut s'attendre à trouver 8 à 18 % de dossiers mal saisis.

Chapitre 3

ESTIMATION PONCTUELLE

☛ *Objectifs*

⇨ Introduire la notion d'estimateur

⇨ étudier ses propriétés

⇨ présenter les méthodes d'estimation ponctuelle

☛ *Mots clés*

Estimateur – variables aléatoires indépendantes et identiquement distribuées – estimateur : convergent ; sans biais ; efficace ; exhaustif – méthode des moments – méthode des moindres carrés – méthode du maximum de vraisemblance – quantité d'information

I *Terminologie ◆ Définitions*

II *Classification des estimateurs*

III *Méthodes d'estimation*

IV *Estimation des paramètres*

AUTO-EVALUATION & EXERCICES CORRIGES

Considérons un ensemble d'observations indépendantes d'une population caractérisée par une variable aléatoire X, ayant une *certaine distribution* connue, dépendant d'*un* ou de *plusieurs paramètres inconnus*.

Exemples de distributions :

* *Distribution de Bernoulli B(1, p)* avec $0 \leq p \leq 1$
* *Distribution Binomiale B(n, p)*
* *Distribution de Poisson P(λ)* avec $\lambda > 0$
* *Distribution normale N(m, σ)* avec *m et* $\sigma > 0$

Nous tirons de cette population, un *échantillon aléatoire* $X_1, \ldots, X_n$, c'est à dire un *échantillon prélevé, au hasard et avec remplacement*, à partir duquel nous cherchons parmi les *fonctions* x_i $f(x_1), \ldots, f(x_n)$, valeurs de X_i celle qui *convient le mieux* pour donner les *meilleures estimations du ou des paramètres inconnus*. Il nous faudra d'abord définir la notion *d'estimateur*, puis donner les *qualités requises* d'une *bonne estimation* et enfin citer les *méthodes* les plus courantes pour l'obtenir.

Nous terminerons le chapitre par *l'estimation des paramètres* les *plus usuels* : *proportion*, *moyenne*, *variance*, ...

I - TERMINOLOGIE • DEFINITIONS

Pour simplifier l'exposé, nous supposons d'abord qu'il existe un *seul paramètre inconnu* θ et nous considérons une *certaine fonction* Θ *ne dépendant pas*, explicitement, de *ce paramètre*.

I - ESTIMATEUR

I - 1. 1 Définitions

On appelle ***estimateur ou statistique*** (en anglais *estimate statistic*), *toute fonction* des observations de l'échantillon appelée *valeurs-échantillon*: $\theta^*(x_1, x_2, \ldots, x_n)$.

I - 1.2 Propriétés

➢ Prélever l'échantillon revient à réaliser *n répétitions indépendantes* d'une expérience aléatoire E: la $i^{ème}$ *répétition* donne lieu à une réalisation x_i de la variable aléatoire **X.**

➢ Les variables aléatoires $X_1, X_2, \ldots, X_n$ sont *n variables aléatoires indépendantes et identiquement distribuées* ***(i.i.d.)***

➢ Comme les X_i sont des variables aléatoires $\Theta = \theta^*(x_1, \ldots, x_n)$, fonction des variables aléatoires $x_1, \ldots, x_n$, est aussi une *variable aléatoire*.

➢ Si plusieurs échantillons de ***n*** valeurs sont prélevés dans la population, et si, pour chacun d'eux, $\Theta = \theta^*(x_1, \ldots, x_n)$, est calculée, *la suite des valeurs obtenues* constitue une suite de *valeurs de la variable aléatoire*.

➢ Dans le cas d'un tirage exhaustif, le *nombre d'échantillons possibles* de taille *n* est égal à : $C_N^n = \dfrac{N!}{n!(N-n)!}$

Remarques

☞ La *distribution de l'estimateur* est appelée ***distribution échantillonnée*** et une ***valeur observée*** de cette variable est appelée ***estimation***.

☞ *L'estimateur* Θ est une formule. C'est la variable aléatoire (***v. a***)

☞ ***L'estimation*** est une ***valeur numérique réelle*** prise par l'estimateur selon les observations (*Ne vous fiez pas à ce chiffre, ce n'est qu'une estimation !).*

☞ Par la suite, nous conviendrons de désigner par $\theta^*(x_1, x_2, \ldots, x_n)$ à la fois *l'estimateur* et *l'estimation*.

☞ Un *estimateur* du paramètre inconnu θ est noté $\hat{\theta}$: θ *estimé*

II - CLASSIFICATION DES ESTIMATEURS

II - 1 ESTIMATEUR CONVERGENT

II - 1.1 Définition

Θ est un estimateur convergent de θ (en anglais : *consistent estimate*) si Θ converge en probabilité vers le paramètre inconnu c'est-à-dire *lim* $\text{Prob}(|\Theta - \theta| > \varepsilon) = 0$, lorsque ***n tend vers l'infini***.

II - 1.2 Exemples

① ***Estimation d'une proportion***

Considérons un lot de *pièces défectueuses* dans la *proportion p*. Attribuons la valeur 1 à la *caractéristique X* lorsque les pièces sont *mauvaises* et la valeur *0* pour les *bonnes pièces*. Il s'agit d'estimer *p* à l'aide d'un échantillon de taille *n*, prélevé au *hasard* avec *remplacement* dans le

lot. La proportion p est égale à la somme des valeurs de la caractéristique divisée par le nombre de pièces du lot.

Si l'on désigne par $\frac{k}{n}$ la fréquence empirique avec laquelle les pièces *défectueuses apparaissent* dans *l'échantillon*, on a:

$$\frac{k}{n} = \overline{X} = \frac{x_1 + \ldots + x_n}{n} = \frac{\sum_1^n x_i}{n} \quad \Rightarrow \quad \frac{k}{n} \text{ est un } \textbf{\textit{estimateur de p}}$$

En utilisant le *théorème de Bernoulli*, on a :

$$\lim \text{Prob}\left(\left| \frac{k}{n} - p \right| > \varepsilon \right) = 0 \quad \textit{lorsque } n \to \infty$$

$\Rightarrow$ l'estimateur $\frac{k}{n}$ *converge en probabilité* vers le *paramètre* ***p***

Pour n très *grand* $\frac{k}{n}$ ***est un estimateur convergent de p.***

Remarque

☞ La *limite pour n tendant vers l'infini* est une *propriété asymptotique*. Elle est insuffisante pour définir un *bon estimateur* car d'autres estimateurs peuvent vérifier cette propriété. A titre d'exemples :

$$\frac{k}{n-1} = \frac{x_1 + \ldots + x_i + \ldots + x_n}{n-1} \quad \text{ou} \quad \frac{k}{n+1} = \frac{x_1 + \ldots + x_i + \ldots + x_n}{n+1}$$

Il faut donc imposer d'autres critères pour le choix des estimateurs.

② ***Généralisation du théorème de Bernoulli***

Soit Θ une statistique, c'est-à-dire une fonction de $x_1, \ldots, x_i, \ldots, x_n$ dont la distribution admet pour moyenne $\overline{\theta}$ et pour écart-type σ tendant vers 0 quand ***n tend vers l'infini***.

D'après *l'inégalité de Bienaymé-Tchebychev*, on a:

$$\text{Prob}\left(\left| \Theta - \overline{\theta} \right| \geq k\sigma \right) \leq \frac{1}{k^2}$$

Comme k est un *nombre positif* quelconque, choisissons-le tel que:
$k\sigma = \varepsilon$ où ε est un *nombre positif voisin de zéro*. On a alors :

$$\text{Prob}\left(\left| \Theta - \overline{\theta} \right| \geq \varepsilon \right) \leq \frac{\sigma^2}{\varepsilon^2} \text{ et } \text{Prob}\left(\left| \Theta - \overline{\theta} \right| \geq \varepsilon \right) = 0 \qquad n \to \infty$$

$\Rightarrow \Theta$ *est un estimateur convergent de* θ.

II - 2 ESTIMATEUR SANS BIAIS

II - 2. 1 Définition

Un estimateur *sans biais* ou *non biaisé* ou *sans distorsion* (en anglais *unbiaised estimate*) est un estimateur dont *l'espérance mathématique* est la *vraie valeur du paramètre à estimer*.

$$E(\hat{\theta}) = \theta$$

L'expression $E(\hat{\theta}) - \theta$ est appelée *biais*

II - 2. 2 Exemples :

① ***Estimation d'une proportion p***

$\frac{k}{n}$ admet une distribution binomiale de paramètre p et d'écart-type

$\sigma = \sqrt{\frac{pq}{n}}$ où $q = 1 - p$; on sait que $E(X) = np$ en conséquence $E(\frac{k}{n}) = p$.

$$\Rightarrow \frac{k}{n} \text{ est un estimateur sans biais de } p.$$

② ***Estimation des paramètres d'une loi normale N(m,σ)***

a) *Estimation de m*

$\bar{x}$ calculée à partir d'un échantillon $(x_1, x_2, \ldots, x_n)$ est *un estimateur non biaisé* de ***m puisque*** $E(\bar{x}) = m$.

⇨ $\bar{x}$ *suit la loi* $N(m, \frac{\sigma}{\sqrt{n}})$.

⇨ $\bar{x}$ est un estimateur *convergent de m.*

Remarques

☞ La population étant *symétrique*, la *moyenne*, la *médiane* et le *mode* sont *égaux* au paramètre *m*. Si *m* est *inconnu*, que faut-il utiliser pour l'estimer ?

Comparons les deux distributions échantillonnées des deux estimateurs. Comme $\bar{x}$ suit la loi $N(m, \frac{\sigma}{\sqrt{n}})$, il est possible de montrer que *Me* suit la loi $N(m, \sqrt{\frac{\pi}{2}}\frac{\sigma}{\sqrt{n}})$ pour ***n grand*** .

☞ Les *deux estimateurs* sont également *sans biais* et *convergents*.

☞ Comme $\sqrt{\frac{\pi}{2}} \# 1{,}25$ *la médiane* $\frac{1}{2}$ possède donc un *écart-type plus grand* que celui de la *moyenne*, sa *distribution* est donc *plus dispersée* et sa fonction de *fréquence plus étalée*. Nous choisirons alors la *moyenne* $\bar{x}$ pour l'***estimation de m***.

b) *Estimation de* σ^2

Soit s^2 *la variance de l'échantillon,* on a :

$$\boxed{E(s^2) = \frac{n-1}{n}\sigma^2}$$

$\mathbf{s^2}$ est un estimateur biaisé de σ^2

Ce biais peut-il être corrigé?

Définissons un *autre estimateur* $S^2 = \frac{ns^2}{n-1}$, on a :

$$E(S^2) = n\frac{E(s^2)}{n-1} = \sigma^2$$

$\mathbf{S^2}$ est donc un *estimateur sans biais* de σ^2

C'est *un estimateur convergent de* σ^2.

Remarque

☞ Lorsqu'on fait tendre *n* vers l'infini le biais disparaît, ***l'estimateur*** est dit *asymptotiquement non biaisé*.

$\lim \frac{n-1}{n} = 1$ lorsque $n \to \infty$ d'où $E(s^2) = \sigma^2$

alors ***s^2est un estimateur asymptotiquement non biaisé de σ^2***

II - 3 ESTIMATEUR EFFICACE

II - 3. 1 Définition :

Un estimateur sans biais est *efficace* ou de *variance minimum* (en anglais *efficient estimate*) si, pour *n* fixé, il est *le plus précis* parmi tous les estimateurs susceptibles d'estimer le paramètre.

Remarques

☞ Pour *n* donné, il est possible de déterminer la *variance minimum* relative à un *paramètre* d'une population.

☞Ayant trouvé un *estimateur sans biais* et connaissant sa *variance*, il suffit alors de *comparer* cette variance avec la *variance minimum*. Si elles sont *égales*, l'estimateur considéré est le meilleur.

☞ Un *estimateur est plus précis* qu'un autre- donc *meilleur* - si pour n fixé, sa *variance est plus petite* que la variance de cet autre estimateur.

☞ Un *estimateur efficace* n'est pas toujours possible à trouver. On utilise alors un *estimateur asymptotiquement efficace*: *sa variance* tend vers la *variance minimum* quand n tend vers *l'infini*.

II - 3. 2 Mesure de l'efficacité

Pour trouver le *meilleur de tous* les estimateurs du *paramètre inconnu* θ, on utilise la *formule de Cramer* appelée également *inégalité de Cramer-Fréchet*

① ***Quantité d'information***

On appelle *quantité d'information de Fisher* fournie par *n-échantillon*, *l'expression suivante*, lorsqu'elle existe :

**** distribution discrète:***

$$I_n(\theta) = E\left[\frac{d\ln P(X=x,\theta)}{d\theta}\right]^2$$

**** distribution continue:***

$$I_n(\theta) = n\,E\left[\frac{\delta \ln f(x,\theta)}{\delta\theta}\right]^2$$

② ***Inégalité de Cramer-Rao***

Il est possible de démontrer que tout estimateur *absolument correct* de θ *vérifie* l'inégalité suivante :

$$V[T_n] \geq \frac{1}{I(\theta)}$$ appelée ***Inégalité de Cramer-Rao***

③ ***Efficacité d'un estimateur***

C'est le *rapport* entre la borne inférieure de l'inégalité de Cramer et la variance de l'estimateur :

$$e = \frac{\frac{1}{I(\theta)}}{Var[T_n]} \quad \text{avec} \quad 0 \leq e \leq 1$$

☞ *L'estimateur* est *efficace* si $V[T_n] = \frac{1}{I(\theta)}$

II - 3. 3 Exemples

❒ Estimation du *paramètre* d'une *loi de Poisson*

① *Quantité d'information*

$$I_n(\lambda) = nE\left(\frac{\delta \ln}{\delta \lambda}(e^{-\lambda}\frac{\lambda^k}{k!}) \right)^2 \Rightarrow I(\lambda) = \frac{n}{\lambda}$$

② *Inégalité de Cramer-Rao*

$$V[T]_n \geq \frac{n}{\lambda}$$

③ Comparaison de deux estimateurs de λ : $\bar{x}$ et S^2

$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n} = \frac{\lambda}{n}$; $\quad e = 1$

$\mathrm{Var}(S^2) = \frac{\lambda}{n} + \frac{2\lambda^2}{n-1}$; $\quad e = \frac{n-1}{n(2\lambda+1)-1}$

$\bar{x}$ est meilleur estimateur que S^2

❒ $\frac{k}{n}$ est un *estimateur efficace* de p.

❒ $\bar{x}$ est un *estimateur efficace de m* dans le cas d'une population normale.

⌧ Par contre *S^2 n'est pas un estimateur efficace* de σ^2 sauf si la population est normale.

II - 4 ESTIMATEUR EXHAUSTIF

II - 4. 1 Définition

Un estimateur ***exhaustif ou suffisant*** (en anglais *sufficient estimate*) est un estimateur qui *traduit toute l'information* relative au paramètre à estimer contenue dans l'échantillon.

II - 4. 2 Propriétés

➢ Si le domaine de variation est indépendant du paramètre, « *une condition nécessaire et suffisante pour que Θ soit un estimateur exhaustif du paramètre θ d'une population discrète (continue) est que la probabilité conditionnelle attachée à l'échantillon, étant donné la valeur estimée $\hat{\theta}$, ne*

dépende pas du paramètre θ pour tout $\hat{\theta}$ où la probabilité (fonction de fréquence) est positive. »

⊠Un estimateur *efficace* est *exhaustif*, mais il existe des estimateurs exhaustifs qui ne *sont pas efficaces*.

II - 4. 3 Exemples

① ***Cas d'une proportion***

Considérons l'exemple 1 celui du lot de pièces défectueuses dans la *proportion p*. A l'aide d'un échantillon de taille ***n***, $(x_1,...,x_n)$ prélevé au *hasard avec remplacement* dans le lot, on estime ***p*** par $\frac{k}{n}$. Peut-on améliorer cette estimation en tenant compte à la fois de la *nature des pièces* considérées dans l'échantillon et de *l'ordre* de *leur apparition*?

Pour *répondre à cette question*, soit :

**A* « *la proportion de pièces défectueuses apparues dans l'échantillon avec une fréquence égale à* $\frac{k}{n}$. »

**B* « *la séquence 110001...01 où 1 est attribué à la mauvaise pièce et 0 à la bonne.* »

La *probabilité p(A)* d'avoir *k* pièces défectueuses dans l'échantillon est fonction de *p inconnu*. Elle est donnée par la *loi binomiale B(n; p)*

$P(A) = C_n^k p^k q^{n-k}$ d'autre part, $P(B) = p^k q^{n-k}$

La probabilité que *B* se réalise sachant que l'événement *A* s'est déjà produit est donnée par *l'axiome des probabilités conditionnelles* :

$$P(B/A) = \frac{P(A \cap B)}{P(B)} = \frac{1}{C_n^k}$$

Cette expression ne dépend pas de p . Ce résultat peut être déterminé en considérant les *évènements équiprobables* : *P(B/A)* désigne la *probabilité* d'avoir un *échantillon déterminé*, sachant qu'il comporte *k pièces défectueuses*.

Nombre de *cas possibles*: C_n^k

Nombre de *cas favorables* : 1 (l'échantillon en question)

D'où $p(B/A) = \frac{1}{C_n^k}$

$\Rightarrow$ $\frac{k}{n}$ ***est un estimateur exhaustif de p***.

② ***Cas d'une moyenne m***

A l'aide d'un échantillon prélevé d'une population normale $N(m, \sigma)$ où *m est inconnu* et σ *connu*, on estime ***m*** au moyen de $\bar{x}$. On ne tient pas compte des valeurs individuelles mais de

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

Peut-on améliorer cette estimation en associant, par exemple, la *médiane* à la *moyenne arithmétique*?

Pour *répondre à cette question*, remplaçons dans l'exemple précédent la *probabilité conditionnelle* par la *densité de probabilité* (la fonction de fréquence) de l'échantillon.

On a :

$$f(x_1,\ldots,x_n/\bar{x}) = \frac{f(x_1,\ldots,x_n)}{f(\bar{x})}$$

En utilisant la *même démarche* que précédemment, on démontre que cette fonction ne dépend pas de *m* et que tout autre information est superflue.

$\Rightarrow$ $\bar{x}$ ***est un estimateur exhaustif de m.***

III - METHODES D' ESTIMATION

Existe-t-il des méthodes générales permettant d'obtenir de ***bons*** *estimateurs* ? Telle est la question posée.

Nous allons passer en revue: la *méthode des moments*, la *méthode des moindres carrés* et la *méthode du maximum de vraisemblance*.

III - 1 METHODE DES MOMENTS

Cette méthode est due à *K.Pearson*. C'est la méthode d'estimation la plus ancienne et la plus simple. Elle consiste à:

- considérer *autant de moments* qu'il y a de *paramètres*
- à *égaler les moments* de l'échantillon- fonctions des paramètres inconnus- *aux moments* correspondants de la population.
- à *résoudre les équations* par rapport à ces *paramètres* pour obtenir les estimations cherchées.

III - 1. 1 Notation

★*Dans la population*

μ_r : le *moment centré* d'ordre r,

$\hat{\mu}_r$: le moment par *rapport à l'origine d'ordre r*

★*Dans l'échantillon*

m_r : le *moment centré* d'ordre r,

$\hat{m}_r$: le moment par rapport à l'origine d'ordre r.

III - 1. 2 Estimation des moments

En appliquant les *propriétés de l'espérance*, on a:

$$E(\hat{m}_r) = E(\frac{1}{n}\sum_{i=1}^{n} x_i^r) = \frac{1}{n}E(\sum_{i=1}^{n} x_i^r) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_r = \hat{\mu}_r$$

En appliquant les *propriétés de la variance*, on a:

$$\mathrm{Var}(\hat{m}_r) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(x_i^2) = \frac{\hat{\mu}_{2r} - \hat{\mu}^2_r}{n}$$

Remarques

☞ $\hat{m}_r$ est un estimateur non biaisé de $\hat{\mu}_r$. C'est un *estimateur convergent.*

⇒ Les *moments centrés* sont donc des *estimateurs convergents* de $\hat{\mu}_r$

☞ Ce sont des estimateurs *biaisés*, mais *asymptotiquement non* ***biaisés***

⇒ Ces *estimateurs* ne sont pas les *meilleurs*.

III - 2 METHODE DES MOINDRES CARRES

On considère un échantillon tiré d'une *population normale de moyenne* μ *inconnue et d'écart type* σ *connu. Comment peut-on estimer* μ ?

La méthode consiste à construire un ***estimateur*** Q qui *minimise l'écart entre les observations et la valeur de la moyenne.*

Pour cela, on définit: $x_i = \mu_i + e_i$ où e_i est une *erreur aléatoire* autour de la *moyenne*. On cherche à *minimiser* l'expression :

$$Q^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(x_i - \mu)^2$$

On calcule alors: $\dfrac{dQ^2}{d\mu} = -2\sum_{i=1}^{n}(x_i - \mu) = 0 \Rightarrow \bar{x} = \dfrac{\sum_{1}^{n} x_i}{n} = \hat{\mu}$

$$\frac{d^2Q}{d\mu^2} = -\frac{d}{d\mu}[(-2\sum_{1}^{n}(x_i-\mu)] > 0$$

Comme la dérivée seconde est positive, alors la *moyenne* de l'échantillon est un *minimum de* Q^2 $\Rightarrow$ $\bar{\mathbf{x}}$ est *un estimateur de* μ.

III - 3 METHODE DU MAXIMUM DE VRAISEMBLANCE MMV ou MV

C'est la méthode la plus importante. Elle a été introduite par ***R.A.Fisher***. Elle consiste à rendre *maximum la probabilité* de prélever l'échantillon observé. Soit $f(x,\theta)$ la *fonction de fréquence* de X dans la *population totale*.

Comme l'échantillon est prélevé *avec remplacement* dans la population, les variables aléatoires- valeurs $x_1,...,x_n$ de X dans l'échantillon- sont *indépendantes*.

➢ Si la ***population*** est ***discrète***

La probabilité de ces n valeurs est donnée par :

$$L(x_1,...x_n,\theta) = p(x_1,\theta)p(x_2,\theta)...p(x_n,\theta)$$

où $p(x_1,\theta)$ *est la probabilité* que x_1 *appartienne* à l'échantillon.

➢ Si la ***population*** est ***continue***

La *densité de probabilité* est donnée par:

$$L(x_1,...x_n,\theta) = f(x_1,\theta)f(x_2,\theta)...f(x_n,\theta)$$

où $f(x_1,\theta)$ est la *fonction de fréquence* de la population en $X = x_1$

III - 3. 1 Définitions

① Cette fonction dépendant des *valeurs de X dans l'échantillon* et du paramètre θ de la *population* qu'on veut estimer est appelée ***fonction de vraisemblance*** ou ***vraisemblance*** (en anglais *likelihood*)

② Θ est un *estimateur maximum de vraisemblance* (en anglais *maximum likehood*) de θ si $\hat{\theta}$ est la valeur de Θ qui rend *maximum* la fonction de vraisemblance. Quand ce *maximum est atteint* dans l'intervalle de variation de $L(\theta)$, la *dérivée de* $L(\theta)$ en ce point *est nulle* :

$$\frac{dL(\theta)}{d\theta} = 0$$

Dans la pratique

☞ On cherche à *maximiser* $\ln L(\theta)$ souvent plus facile à calculer. Ce maximum est le même que celui de L(θ) puisque la *fonction logarithme* est une *fonction monotone croissante*.

☞ On annule la *dérivée* : $\frac{d\ln L(\theta)}{d\theta} = 0$

☞ Cette équation est ***l'équation du maximum de vraisemblance***

III - 3. 2 Propriétés de la MMV:

La méthode du maximum de vraisemblance, a les propriétés suivantes que nous admettrons :

① *S'il existe* un *estimateur efficace, il est la solution* de l'équation du *maximum de vraisemblance*.

② S'il existe un *estimateur exhaustif, toute solution* de l'équation du maximum de vraisemblance est *fonction de cet estimateur*.

③ Si l'équation du maximum de vraisemblance admet une solution unique alors :

✓ Θ est un estimateur *convergent* et *asymptotiquement efficace* du paramètre θ

✓ Θ admet une distribution *asymptotiquement normale*.

III - 3. 3 Applications de la MMV

① ***Distribution de Bernoulli B(1; p)***

Considérons une population pour laquelle la caractéristique X admet une ***distribution de Bernoulli*** de *paramètre* ***p*** *inconnu*.

On considère une expérience aléatoire E et un évènement A de probabilité p *inconnu*. On effectue *une suite de n répétitions*. Soit r le nombre de réalisations de l'évènement A (dans ce cas la valeur de $X = 1$, dans le cas contraire $X = 0$).

Les n observations constituent un échantillon aléatoire et r est le nombre d'unités figurant dans la suite. La *fonction de vraisemblance* est donnée par:

$$L(p)=p^r q^{n-r}$$

Comment estimer p?

•Calculons *lnL(p)*

$$\ln L(p) = r\ln p - (n-r)\ln q$$

•Dérivons par rapport à **p** et annulons la dérivée :

$$\frac{d\ln L(p)}{dp} = \frac{r}{p} - \frac{n-r}{1-p} = 0 \Rightarrow p = \frac{r}{n}$$

☞ $\frac{r}{n}$ est une distribution binomiale de paramètre p et d'exposant n.

☞ $\frac{r}{n}$ est un estimateur absolument correct et convergent de p

☞ $\frac{r}{n}$ est un estimateur asymptotiquement efficace.

② ***Distribution de Poisson***

Considérons une population pour laquelle la caractéristique X admet une ***distribution de Poisson*** de paramètre λ *inconnu* :

$$P(X=x) = \frac{e^{-\lambda}m^{x}}{x!} \qquad \textit{Comment estimer } \lambda \textit{ ?}$$

Appliquons la méthode du *maximum de vraisemblance* sur un échantillon d'effectif n : $x_1,\ldots, x_n$

$$L(x_1,\ldots x_n, \lambda) = p(x_1, \lambda)\times p(x_2, \lambda)\times\ldots\times p(x_n, \lambda)$$

$$L(\lambda) = \frac{e^{-\lambda}m^{x1}}{x_1!}\times\frac{e^{-\lambda}m^{x2}}{x_2!}\times\ldots\times\frac{e^{-\lambda}m^{xn}}{x_n!} \qquad \text{et} \quad L(\lambda) = \frac{e^{-n\lambda}m_1^{\sum x_i}}{x_1!\ldots x_n!}$$

•Calculons lnL(λ)

$$\ln L(\lambda) = -n\lambda + \sum_1^n x_i \ln\lambda - \sum_1^n \ln x_i!$$

•Dérivons par rapport à λ et annulons la dérivée:

$$\frac{d\ln L(\lambda)}{d\lambda} = -n + \sum_1^n x_i \frac{1}{\lambda} = 0 \Rightarrow \lambda = \frac{\sum_1^n x_i}{n} = \bar{x}$$

☞ $\bar{x}$ est un estimateur *convergent* et *asymptotiquement efficace* de λ.

☞ $\lambda = \bar{x}$ est un *estimateur efficace*.

☞ Dans une population qui suit la loi de Poisson, on estime le paramètre λ par la *moyenne arithmétique*.

☞ Rappelons que la *variance* d'une loi de Poisson est aussi égale à λ.

③ **Distribution normale**

Considérons une population pour laquelle la caractéristique X admet une ***distribution normale*** $N(m, \sigma)$ de paramètre m *et* σ*; de densité de probabilité*:

$$f(m;\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Trois cas à étudier:

a) *m inconnue et σ connu*

Il s'agit *d'estimer m* à partir d'un échantillon de taille n : $x_1,\ldots x_n$

x_i admet une distribution $N(m\ ;\ \sigma)$, D'où

$$f(x_i;m)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-m)^2}{2\sigma^2}}$$

La densité de probabilité est donnée par:

$$L(x_1,\ldots x_n,\theta)=f(x_1,\theta)f(x_2,\theta)\ldots f(x_n,\theta)$$

$$L(x_1,\ldots,x_i\ldots,x_n;m)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_1-m)^2}{2\sigma^2}}\times\ldots\times\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_n-m)^2}{2\sigma^2}}$$

$$=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\sum_1^n(x_i-m)^2}{2\sigma^2}}$$

•Calculons $\ln L$

$$\ln L(x_1,\ldots x_n,m)=n\ln\frac{1}{\sigma\sqrt{2\pi}}-\frac{1}{2\sigma^2}\sum_1^n(x_i-m)^2$$

•*Dérivons par rapport à* ***m*** et *annulons la dérivée*:

$$\frac{d\ln L(m)}{dm}=\frac{1}{\sigma^2}\sum_1^n(x_i-m)=0\Rightarrow m=\frac{\sum_1^n x_i}{n}=\bar{x}$$

b) *m connue et σ inconnu*

$$L(x_1,\ldots x_n,\sigma^2)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\sum_1^n(x_i-m)^2}{2\sigma^2}}$$

•Calculons $\ln L$

$$\ln L(x_1,\ldots x_n,m)=n\ln\frac{1}{\sqrt{2\pi}}-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_1^n(x_i-m)^2$$

•Dérivons par rapport à σ^2 et annulons la dérivée:

$$\frac{d\ln L}{d\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_1^n (x_i - m)^2 = 0 \Rightarrow \sigma^2 = \frac{1}{n}\sum_1^n (x_i - m)^2$$

c) *m inconnue et σ inconnu*

Pour trouver le maximum, il faut alors ***annuler les dérivées partielles par rapport aux deux paramètres.***

$$L(x_1,..,x_i..,x_n; m, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\sum_1^n (x_i - m)^2}{2\sigma^2}}$$

• *Calculons lnL*

$$\ln L(x_1,...x_n;\ m, \sigma^2) = n\ln\frac{1}{\sqrt{2\pi}} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_1^n (x_i - m)^2$$

• *Dérivons par rapport à* ***m*** et annulons la dérivée:

$$\frac{d\ln L(m)}{dm} = \frac{1}{\sigma^2}\sum_1^n (x_i - m) = 0 \Rightarrow m = \frac{\sum_1^n x_i}{n} = \bar{x}$$

• *Dérivons par rapport à* σ^2 et annulons la dérivée:

En utilisant $\boldsymbol{m} = \bar{x}$

$$\frac{d\ln L}{d\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_1^n (x_i - \bar{x})^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$$

$$\Rightarrow \textbf{\textit{C'est la variance } s^2}$$

☞ Les deux estimateurs trouvés par la *MMV* sont ***convergents*** et ***asymptotiquement efficaces***.

En résumé

Pour trouver l'estimateur du maximum de vraisemblance d'un paramètre θ, on procède comme suit:

① ***Ecrire explicitement la fonction de vraisemblance***

$$L(x_1,...x_n, \theta) = f(x_1, \theta)f(x_2, \theta)...f(x_n, \theta)$$

② ***Prendre le logarithme de la fonction de vraisemblance***

$$\boldsymbol{logL} = \prod_1^n f(x_i, \theta)$$

③ ***Dériver le logarithme par rapport θ et calculer la valeur pour laquelle la dérivée est nulle***

④ ***Calculer la dérivée seconde au point trouvé. Si cette dérivée est négative*** $\Rightarrow \Theta$ ***est l'estimateur du maximum de vraisemblance.***

Remarque

☞ Si la *distribution de X* dans la population totale est définie par une ***fonction de fréquence*** dépendant de ***plusieurs paramètres*** qu'on désire *estimer simultanément*, on procède comme suit:

❶ *Ecrire la vraisemblance de l'échantillon qui dépend de ces paramètres* ;
❷ *Appliquer la MMV comme pour le cas de* $\mathbf{m}$ *et* σ^2 *inconnus* ;
❸ *Calculer les dérivées partielles pour les différents paramètres* ;
❹ *Annuler ces dérivées* $\Rightarrow$ *on obtient alors un système d'équations* ;
❺ *Les estimations seront déterminées par les solutions (si elles existent) de ce système.*

IV - ESTIMATION DES PARAMETRES

IV - 1 ESTIMATION D'UNE PROPORTION

IV - 1. 1 Notation

On considère une population partagée en deux classes suivant que ses éléments *possèdent ou non* un caractère. Le caractère X peut donc prendre *deux modalités* désignées par A et $\overline{A}$. On associe la valeur *1* à la modalité A et *0* à la modalité $\overline{A}$. En supposant que tous les éléments aient la même probabilité d'être choisis, on réalise l'*expérience aléatoire* qui consiste à choisir au hasard un élément de la population.

Le caractère X apparaît comme une *variable aléatoire* qui prend les valeurs *1* et *0* avec les *probabilités* p et $1 - p = q$, où p désigne la *proportion* dans laquelle les *éléments* de la population *possèdent le caractère* A. Cette variable admet une distribution de *Bernoulli B(1, p)*. Supposons qu'on ne peut pas effectuer un *recensement,* la proportion p est alors *inconnue* et on cherche à *l'estimer*.

Si l'on désigne par $(x_1, ..., x_n)$ un échantillon de taille n, prélevé au hasard avec remplacement dans cette population $\boldsymbol{P}$, ces valeurs-échantillons sont les valeurs observées de n variables aléatoires $x_1,..., x_n$ *indépendantes* et *identiquement distribuées* ***(i.i.d)*** de même distribution.

IV- 1. 2 Estimation de *p*

Si l'on désigne par $\frac{k}{n}$ la *fréquence relative* avec laquelle la ***modalité*** *A* apparaît dans l'échantillon, on a :

$$\frac{k}{n} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_1^n x_i$$

Cette *fréquence* peut être considérée comme la valeur observée d'une variable aléatoire, moyenne de *n variables aléatoires indépendantes* de *Bernoulli*

$$\Rightarrow \quad \frac{k}{n} \text{ est un } \textbf{\textit{estimateur de p}}$$

IV- 1. 3 Propriétés de $\frac{k}{n}$

Les *caractéristiques* de *l'estimateur de p* sont :

- ✓ $\frac{k}{n}$ admet une distribution *binomiale B(n, p)*
- ✓ $\frac{k}{n}$ est un *estimateur convergent* de *p*
- ✓ $E(\frac{k}{n}) = p \Rightarrow \frac{k}{n}$ est un *estimateur sans biais* de *p*
- ✓ $V(\frac{k}{n}) = \frac{pq}{n} \Rightarrow \frac{k}{n}$ est un *estimateur efficace* de *p*
- ✓ $\frac{k}{n}$ est la *solution unique* de l'équation du ***maximum de vraisemblance***.

$$\Rightarrow \quad \frac{k}{n} \textbf{\textit{ est un estimateur convergent et efficace}}.$$

IV - 2 <u>ESTIMATION D'UNE MOYENNE</u>

Soit une population ***P*** pour laquelle on s'intéresse au caractère *X*. Soit *m sa moyenne* et σ *son écart-type*. On y prélève un *échantillon aléatoire* de *taille n*.

IV - 2. 1 <u>Notation</u>

- $\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_1^n x_i$ (ou $\overline{X}$) la *moyenne* de l'échantillon
- $s^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$ la *variance* de l'échantillon

IV - 2. 2 Estimation de $\bar{x}$

Deux cas à examiner :

❶ ***Population quelconque*** :

échantillon de taille n assez grand (n> 30)

$\Rightarrow E(\bar{x}) = m \Rightarrow \text{Var}(\bar{x}) = \frac{\sigma^2}{n}$

$\Rightarrow \bar{x} \rightarrow N(m, \frac{\sigma}{\sqrt{n}})$ *est une distribution approximativement normale.*

❷ ***Population normale :***

✓ $\bar{x} \rightarrow N(m, \frac{\sigma}{\sqrt{n}})$

c'est une distribution normale quelle que soit la taille de l'échantillon.

➢ *Les propriétés précédentes de $\bar{x}$ demeurent valables*

IV - 2. 3 Propriétés de $\bar{x}$

✓ *$\bar{x}$ est un estimateur sans biais de* ***m***

✓ *$\bar{x}$ est un estimateur convergent de* ***m***

$\Rightarrow$ ***$\bar{x}$ est un estimateur convergent et efficace de m***

IV - 3 <u>ESTIMATION DE LA VARIANCE</u>

IV - 3. 1 <u>Notation</u>

- $S^2 = \frac{ns^2}{n-1}$ *l'estimation non biaisée* ou *quasi-variance* ou *écart quadratique*
- μ_r *le moment centré d'ordre r*

IV - 3. 2 Estimation de σ^2

✓ $E(s^2) = \frac{n-1}{n}\sigma^2 \Rightarrow s^2$ est un estimateur biaisé de σ^2

✓ $E(S^2) = \sigma^2 \Rightarrow S^2$ est un estimateur sans biais de σ^2

Remarques

☞ L'emploi de S^2 est justifiée parce que c'est un estimateur sans biais de σ^2

☞ Dans le cas où $n \geq 50$, la distinction entre s^2 et S^2 ne serait *pas importante.*

Tableau - synthèse des estimateurs par la MMV

Nom de la loi	*Densité de probabilité : f(x, θ)*	*Paramètre à estimer:* θ	*MMV*
Bernoulli	$p^x(1-p)^{1-x}$	p	$\hat{\theta} = \overline{X}$
Binomiale	$C_n^x p^x (1-p)^{n-x}$	p	$\hat{\theta} = \frac{\overline{X}}{k}$
Géométrique	$p(1-p)^{x-1}$	p	$\hat{\theta} = \frac{1}{\overline{X}}$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ	$\hat{\theta} = \overline{X}$
Normale	$\frac{1}{\sigma\sqrt{\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	μ *σ connu*	$\hat{\theta} = \overline{X}$
Normale	$\frac{1}{\sigma\sqrt{\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	σ *μ connue*	$\hat{\theta} = \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{n}$
Normale	$\frac{1}{\sigma\sqrt{\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	σ μ	$\hat{\mu} = \overline{X}$ $\hat{\theta} = \sum_{i=1}^{n} \frac{(X_i-\overline{X})^2}{n}$
Exponentielle	$k e^{-kx}$	k	$\hat{\theta} = \frac{1}{\overline{X}}$

Auto-évaluation ?

1- La moyenne empirique est un estimateur sans biais de la moyenne arithmétique :

a) Vrai b) Faux

2- La variance est un estimateur :

a) Sans biais b) biaisé c) convergent

3- Etant donné un échantillon de taille n d'une v.a. X suivant une loi de Bernoulli de paramètre p inconnu :

a) $\hat{p} = \bar{x}$ b) $\hat{p}$ est convergent

c) $\hat{p}$ est efficace d) $E(\hat{p}) = \dfrac{p(1-p)}{n}$

4- Soit n v. a. indépendantes et de même loi, déterminer les lois de

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{et} \qquad Y = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

5- Un échantillon de taille $n = 25$ et $\sigma = 2$ déterminer a telle que :

$$\text{Prob}\left(\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{25} > a\right) = 0{,}05$$

6- Une variable aléatoire X suit une loi normale d'espérance mathématique m et de variance σ^2. Soit $(X_1, \ldots, X_{n-1})$ un échantillon de la variable X

① Déterminer les lois de

$$\frac{\sum_{i=1}^{n-1} X_i}{n-1} \quad ; \qquad \sum_{i=1}^{n-1} \frac{(X_i - \bar{X})^2}{\sigma^2} \quad ; \qquad (n-1)\frac{(\bar{X} - m)^2}{\sigma^2}$$

② On suppose que ses deux paramètres sont inconnus. Donner des estimateurs sans biais de m et de σ^2.

7- Soit $(X_1, .., X_n)$ un échantillon de taille n de la variable X qui admet pour espérance mathématique $E(X) = \theta$ et pour variance $V(X) = \sigma^2$. θ est un paramètre inconnu. On définit les estimateurs suivants :

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} X_i}{n} \quad ; \quad \hat{\theta}_2 = \frac{1}{2}X_1 + \frac{X_2 + \ldots + X_{n-1}}{n-2} + \frac{1}{2}X_n$$

Ces estimateurs sont-ils sans biais ? Peut-on les comparer ?

8- Soit (X_1, X_2) un échantillon de taille $n = 2$ de la variable X. Exprimer en fonction de X l'estimation de l'écart-type.

9- Soit $(X_1, \ldots, X_n)$ un échantillon de taille n de la variable X qui admet pour espérance mathématique $E(X) = m$ et pour variance $V(X) = \sigma^2$

① Déterminer $E(\overline{X})$ et $V(\overline{X})$

② Déterminer la loi de $Z = \sum_{i=1}^{n} \frac{(X_i - m)^2}{\sigma^2}$

③ Déterminer $E(Z)$ et $V(Z)$

10- Soit $(X_1, .., X_n)$ un échantillon de taille n de la variable X qui admet pour espérance mathématique $E(X) = m$ et pour variance $V(X) = \sigma^2$

① Déterminer la loi de $\overline{X}$

② On définit la quasi-variance : $S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$

Déterminer la loi de $Z = (n-1)\frac{S^2}{\sigma^2}$

Exercices ?

Exercice 1

Soit x une variable continue dont la densité dépend d'un paramètre $\theta > 0$

$$f(x,\theta) = \frac{3x^2}{\theta}\exp(-\frac{x^3}{\theta}) \text{ avec } x > 0$$

1- Soit un échantillon de taille n extrait d'une variable aléatoire, de même loi que X. Déterminer l'estimateur du maximum de vraisemblance de Θ, notée $\hat{\theta}$.

2- Calculer le score $S(x, \theta)$ et en déduire $E(X^3)$

3- Calculer la quantité d'information de Fisher $I(\theta)$ et en déduire $V(X^3)$

4- Montrer que Θ est un estimateur sans biais et convergent de θ. Est-il efficace ?

5- On considère la variable $Y = \frac{x^3}{\theta}$. Calculer la densité de Y. Montrer que Y suit une loi usuelle (à préciser) et en déduire $E(X^3)$ et $V(X^3)$

N. B. $\qquad S(x,\theta) = \frac{\delta \log f(x,\theta)}{\delta\theta}$

D'après examen février 1996, ISG Tunis.

Exercice 2

On considère la répartition des ménages tunisiens en deux classes C_1 et C_2 suivant qu'ils habitent la ville ou la campagne. On désigne respectivement par p_1 et p_2 les proportions dans chaque classe des ménages possédant une parabole. Pour estimer p_1 et p_2, on a tiré au hasard et avec remise n_1 ménages dans C_1 et n_2 ménages dans C_2 et pris pour estimateurs les fréquences F_1 et F_2 respectives des ménages possédant une parabole dans chaque échantillon. On suppose alors $p_1 = p_2 = p$

1- Les estimateurs F_1 et F_2 et $\frac{F_1 + F_2}{2}$ sont-ils sans biais ?

2- On suppose $n_1 > n_2$ quel est le meilleur des estimateurs précédents ?

3- Parmi les estimateurs de la forme $aF_1 + bF_2$ quel est le meilleur estimateur sans biais.

Exercice 3

On désigne par X la durée d'une communication téléphonique. On admet que X suit une loi uniforme sur $[0 ; \theta]$. On observe les durées de n communications.

1- Calculer l'espérance E(X). En déduire un estimateur de θ. Cet estimateur $\hat{\theta}_1$ est-il sans biais ? convergent ?

2- Déterminer l'estimateur du maximum de vraisemblance de Θ noté $\hat{\theta}_2$. Est-il sans biais ? convergent ?

3- On définit l'estimateur sans biais et convergent $\hat{\theta}_3$ proportionnel à $\hat{\theta}_2$. Comparer $\hat{\theta}_1$ et $\hat{\theta}_3$

4- Calculer la quantité d'information de Fisher $I(\theta)$ et la comparer aux résultats de (1).

Exercice 4

Soit X une variable aléatoire représentant la durée d'attente dans un service hospitalier dans un hôpital A et admettant une densité de probabilité :

$$f(x, \theta) = \begin{cases} \frac{1}{\theta} e^{\frac{-x}{\theta}} & \forall\, x > 0 \\ 0 & \text{sinon} \end{cases}$$

où θ est un paramètre inconnu. On prélève un échantillon de taille n.

1- Calculer l'espérance mathématique de X et sa variance.

2- Montrer que Θ est un estimateur sans biais et convergent de θ. Est-il efficace ?

Exercice 5

Pour connaître la proportion p des services hospitaliers possédant un ordinateur au moins, un organisme d'étude effectue un sondage. L'enquête montre que $n = 100$ services choisis au hasard, 25 possèdent un ordinateur au moins.

1- Donner une estimation ponctuelle $\hat{p}$ de p.

2- Montrer que $\hat{p}$ est un estimateur sans biais et convergent.

3- Déterminer la distribution de la fréquence $F_n = \hat{p}$ et calculer ses caractéristiques.

Corrigés

Auto-évaluation

1- Vrai

2- biaisé; convergent

3- $\hat{p} = \bar{x}$; $\hat{p}$ est convergent ; $\hat{p}$ est efficace

4- $\bar{X} \to N(m, \frac{\sigma}{\sqrt{n}})$ et de $Y \to \chi^2(n-1)$

5- $\text{Prob}\left(\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} > \frac{25a}{\sigma^2}\right) = 0{,}05$; $\frac{25a}{\sigma^2} = 36{,}41 \Rightarrow$ a= 5,8286 # 5,83

6- ①

- $\frac{\sum_{i=1}^{n-1} X_i}{n-1} \to N(m, \frac{\sigma^2}{n-1})$
- $\sum_{i=1}^{n-1} \frac{(X_i - \bar{X})^2}{\sigma^2} \to \chi^2(n-2)$
- $(n-1)\frac{(\bar{X} - m)^2}{\sigma^2} \to \chi^2(1)$

② Estimateurs sans biais de m et de σ^2

a) $\hat{m} = \bar{X}$ b) $\hat{\sigma}^2 = \sum_{i=1}^{n-2} \frac{(X_i - \bar{X})^2}{n-2}$

7-

$E(\hat{\theta}_1) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = E(X_i) = \theta \Rightarrow \hat{\theta}_1$ est un estimateur sans biais de Θ.

$E(\hat{\theta}_2) = \frac{1}{2}E(X_1) + \frac{E(X_2) + \ldots + E(X_{n-1})}{n-2} + \frac{1}{2}E(X_n) = 2\theta$

$\Rightarrow \hat{\theta}_2$ est un estimateur biaisé de Θ.

On ne peut pas les comparer puisque *l'un est sans biais* et *l'autre biaisé*.

8- Soit x_1 *et* x_2 les valeurs de X, on a :

$$\bar{x} = \frac{x_1 + x_2}{2} \text{ puis } S^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2}{n-1} \text{ avec } n = 2$$

$$S^2 = \frac{(x_1 - x_2)^2}{2} \qquad \text{d'où} \quad S = \frac{|x_1 - x_2|}{\sqrt{2}}$$

$\Rightarrow$ *C'est un estimateur sans biais.*

9- ①

$$E(\bar{X}) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{1}{n} E(\sum_{i=1}^{n} X_i) = \frac{nm}{n} = m$$

$$V(\bar{X}) = \frac{\sum_{i=1}^{n} V(X_i)}{n} = \frac{1}{n^2} V(\sum_{i=1}^{n} X_i) = \frac{\sigma^2}{n}$$

② $Z \rightarrow \chi^2(n-1)$ C'est une somme de n *v. a.* indépendantes suivant chacune une loi normale centrée réduite

③ $E(Z) = n$ $\qquad$ $V(Z) = 2n$

10-

① $\bar{X} \rightarrow N(m, \frac{\sigma}{\sqrt{n}})$

② $Z \rightarrow \chi^2(n-1)$

Exercises

Exercice 1

① *Détermination de l'estimateur*

La vraisemblance s'écrit :

$$L(x_1,\dots x_n, \theta) = 3^n\, \theta^{-n} \prod_{i=1}^{n} x^2{}_i \exp(-\frac{1}{\theta}) \sum_{i=1}^{n} x_i^3$$

En passant au logarithme, on a :

$$\ln L(x_1,\dots x_n, \theta) = n \ln 3 - n\ln\theta + 2\ln \prod_{i=1}^{n} x_i - \frac{1}{\theta} \sum_{i=1}^{n} x_i^3$$

On dérive par rapport à θ on obtient :

$$\frac{\delta \ln L}{\delta\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i^3 \qquad \frac{\delta^2 \ln L}{\delta\theta^2} = -\frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^{n} x_i^3$$

La dérivée première s'annule pour $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i^3$.

En ce point la dérivée seconde est négative.

L'estimateur est donc $\Theta = \frac{1}{n} \sum_{i=1}^{n} x_i^3$

② *Calcul du score*

$$S(x,\theta) = \frac{\delta \ln f(x,\theta)}{\delta\theta} \text{ d'où } S(x, \theta) = -\frac{1}{\theta} + \frac{x^3}{\theta^2}$$

et $E[S(x, \theta)] = 0 \Rightarrow E(x^3) = \theta$

③ *Calcul de $I(\theta)$*

$$I(\theta) = E[S^2(x, \theta)] = -E(\frac{\delta S}{\delta\theta}) \Rightarrow I(\theta) = \frac{1}{\theta^2}$$

$$I(\theta) = V[S(x,\theta)] = \frac{1}{\theta^4} V(x^3) = \frac{1}{\theta^2} \Rightarrow V(x^3) = \theta^2$$

④ *Propriétés de $\hat{\theta}$*

$$E(\hat{\theta}) = \theta \Rightarrow \hat{\theta} \textit{ est sans biais et } V(\hat{\theta}) = \frac{\theta^2}{n} \to 0 \Rightarrow \hat{\theta} \text{ est } \textit{convergent}$$

$$I_n(\theta) = nI(\theta) = \frac{1}{V(\hat{\theta})} \Rightarrow \hat{\theta} \text{ est } \textit{efficace}$$

⑤ *Calcul de la densité de Y*

Soit $G(y)$ la *fonction de répartition* de Y définie par :

$G(y) = \text{Prob}(Y < y) = \text{Prob}(Y < y) = \text{Prob}(\frac{X^3}{\theta} < y) = \text{Prob}(X < \theta^3 y^3)$

En dérivant, on obtient : $g(y) = e^{-y}$.

Y est une *loi exponentielle* de paramètre 1 dont les caractéristiques sont :

$$E(Y) = V(Y) = 1$$

$$E(\frac{X^3}{\theta}) = 1 \Rightarrow E(X^3) = \theta \quad \text{et} \quad V(\frac{X^3}{\theta}) = \frac{1}{\theta^2} V(X^3) \Rightarrow V(X^3) = \theta^2.$$

Exercice 2

① *Propriétés des trois estimateurs*

$$E(F_1) = p_1 = p \; ; \; E(F_2) = p_2 = p \text{ et } E(\frac{F_1 + F_2}{2}) = \frac{p_1 + p_2}{2} = p$$

$\Rightarrow$ *Les estimateurs sont sans biais*

② *Recherche du meilleur estimateur*

$$V(F_1) = \frac{pq}{n_1} \; ; V(F_2) = \frac{pq}{n_2}$$

Comme F_1 *et* F_2 sont *indépendants* :

$$V(\frac{F_1 + F_2}{2}) = \frac{V(F_1) + V(F_2)}{4} = \frac{pq}{4}(\frac{1}{n_1} + \frac{1}{n_2})$$

* $n_1 > n_2$ alors $V(F_1) < V(F_2)$, on en déduit que F_1 est *meilleur* que F_2

$$\frac{V(\frac{F_1 + F_2}{2})}{V(F_2)} = \frac{pq}{4}(\frac{1}{n_1} + \frac{1}{n_2})\frac{n_2}{pq} < 1 \quad \text{si} \quad n_2 < 3n_1$$

$$\Rightarrow \frac{F_1 + F_2}{2} \text{ est } \textit{meilleur} \text{ que } F_2$$

Le même raisonnement nous conduit à

$$V(\frac{F_1 + F_2}{2}) < V(F_1) \quad \text{si} \quad n_1 < 3n_2$$

$\Rightarrow$ F_1 est *meilleur* que $\frac{F_1 + F_2}{2}$ *meilleur* que F_2

③ *Recherche du meilleur estimateur sans biais*

$$E(aF_1 + bF_2) = (a + b)p \quad \text{avec } (a + b) = 1$$

$$V(aF_1 + bF_2) = a^2 V(F_1) + b^2 V(F_2) = a^2 \frac{pq}{n_1} + b^2 \frac{pq}{n_2}$$

Il s'agit de *déterminer a et b* pour que la *variance soit minimale*.

$$\begin{cases} a + b = 1 \\ \dfrac{a^2}{n_1} + \dfrac{b^2}{n_2} \qquad \textit{minimum} \end{cases}$$

Pour *résoudre ce système*, on utilise la *méthode de Lagrange*.

On définit alors une *fonction* $h(a, b, \lambda) = \dfrac{a^2}{n_1} + \dfrac{b^2}{n_2} + \lambda(a + b - 1)$ et on *annule ses dérivées* par *rapport à a et à b*, on obtient après résolution les valeurs :

$$a = \frac{n_1}{n_1 + n_2} \qquad \text{et} \qquad b = \frac{n_2}{n_1 + n_2}$$

$\Rightarrow$ l'estimateur sans biais de variance minimum est $\hat{P} = \dfrac{n_1 F_1 + n_2 F_2}{n_1 + n_2}$.

Cette valeur correspond à la *fréquence dans l'échantillon global*.

Exercice 3

① *Calcul de E(X)*

La densité de probabilité de X est donnée par

$$f(x, \theta) = \begin{cases} \dfrac{1}{\theta} & \text{si } 0 \le x \le \theta \\ 0 & \textit{sinon} \end{cases}$$

donc $$E(X) = \int_0^{\theta} \frac{x}{\theta}\, dx = \frac{\theta^2}{2\theta} \quad \Rightarrow E(X) = \frac{\theta}{2}$$

On sait que la *moyenne* est toujours *un estimateur sans biais et convergent* de *E(X)*, donc

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ est } \textit{un estimateur sans biais et convergent de } \frac{\theta}{2}$$

$$\Rightarrow \hat{\theta}_1 = 2\frac{\sum_{i=1}^{n} X_i}{n} \text{ est } \textit{un estimateur sans biais et convergent de } \theta$$

② *Détermination de l'estimateur par la MMV*

La vraisemblance s'écrit :

$$L(x_1, \ldots x_n, \theta) = \begin{cases} \theta^{-n} & 0 \le x_i \le \theta \\ 0 & \textit{sinon} \end{cases}$$

$L(x_1,\ldots,x_n,\theta)$ sera *maximum* pour θ *minimum*.

$\hat{\theta}_2$ est le plus petit des majorants des $x_i \Rightarrow \hat{\theta}_2 = sup(X_1, X_2,\ldots, X_n)$

Détermination de la loi de $\hat{\theta}_2$

**Fonction de répartition* :

$G(x) = [F(x)]^n$ ou F(x) est la *fonction de répartition* de X définie par :

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{x}{\theta} & \text{si } 0 \le x \le \theta \\ 1 & \text{si } x > \theta \end{cases} \qquad G(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{x^n}{\theta^n} & \text{si } 0 \le x \le \theta \\ 1 & \text{si } x > \theta \end{cases}$$

* *Densité de probabilité*

$$g(x) = \begin{cases} \frac{nx^{n-1}}{\theta^n} & \text{si } 0 \le x \le \theta \\ 0 & \text{sinon} \end{cases}$$

* *Propriétés de l'estimateur*

$E(\hat{\theta}_2) = \frac{n}{n+1}\theta \Rightarrow$ *il est biaisé* $V(\hat{\theta}_2) = \frac{n}{(n+2)(n+1)^2}\theta^2$

quand *n tend vers l'infini* $E(\hat{\theta}_2) \to \theta$ et $V(\hat{\theta}_2) \to \theta$ donc $\hat{\theta}_2$ est *convergent*

③ *Définition de* $\hat{\theta}_3$

On définit

$$\hat{\theta}_3 = \frac{n+1}{n}\hat{\theta}_2 \Rightarrow \hat{\theta}_3 \text{ est sans biais}$$

$$V(\hat{\theta}_3) = V(\frac{n+1}{n}\hat{\theta}_2) = \frac{\theta^2}{n(n+2)} \Rightarrow \hat{\theta}_3 \text{ est convergent}$$

Comparaison des deux estimateurs

$$\frac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)} = \frac{n(n+2)}{3n} > 1 \quad \text{si } n > 1 \qquad \Rightarrow \hat{\theta}_3 \text{ est } meilleur \text{ que } \hat{\theta}_1$$

④ *Calcul de la quantité d'information*

$$I(\theta) = E[(\frac{\partial \ln L}{\partial \theta})^2]$$

En passant au logarithme, on a : $\ln L = \sum_{i=1}^{n} \ln\frac{1}{\theta} = -n\ln\theta$ et en dérivant

$$\frac{\delta \ln L}{\delta\theta} = -\frac{n}{\theta} \quad \Rightarrow I(\theta) = \frac{n^2}{\theta^2} \text{ et } V(\hat{\theta}_3) = \frac{\theta^2}{n(n+2)} < \frac{1}{I(\theta)}$$

L'inégalité de Cramer-Rao n'est pas *vérifiée* car ses *conditions d'application ne sont pas satisfaites* puisque la *densité de probabilité* de X dépend du paramètre θ.

Exercice 4

① *Calcul des caractéristiques*

❶ *Calcul de E(X)*

$E(X) = \int_0^{+\infty} \frac{x}{\theta} e^{\frac{-x}{\theta}} dx$ on intègre par partie, en *posant* $x = u$ et $dx = dv$

On obtient $E(X) = [-x e^{-\frac{x}{\theta}}]_0^{+\infty} - \int_0^{+\infty} -e^{-\frac{x}{\theta}} dx \quad \Rightarrow E(X) = \theta.$

❷ *Calcul de V(X)*

$V(X) = E(X^2) - [E(X)]^2 \quad * \quad E(X^2) = \int_0^{+\infty} \frac{x^2}{\theta} e^{\frac{-x}{\theta}} dx$

on intègre par partie, en *posant* $x = u$ et $dv = \frac{1}{\theta} e^{-\frac{x}{\theta}} dx$

On obtient : $E(X^2) = [-x^2 e^{-\frac{x}{\theta}}]_0^{+\infty} - \int_0^{+\infty} -2xe^{-\frac{x}{\theta}} dx = 2\theta^2 \Rightarrow V(X) = \theta^2.$

② *Détermination de l'estimateur par la MMV*

$$L(x_1,\ldots x_n,\theta) = \frac{1}{\theta^n} e^{-\frac{\sum_i x_i}{\theta}} \Leftrightarrow \ln L = -n\log\theta - \frac{\sum_1^n x_i}{\theta}$$

En dérivant par rapport à θ : $\frac{\delta \ln L}{\delta\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i \quad \Rightarrow \hat{\theta} = \bar{x}$

③ *Propriétés de* θ

$E(\bar{X}) = E(F_n) = p \Rightarrow \hat{p}$ *est sans biais* ; $V(\bar{X}) = V(\frac{\sum_{i=1}^{n}(X_i)}{n}) = \frac{p(1-p)}{n}$

Or $\lim_{n\to\infty} V(\bar{X}) = \lim_{n\to\infty} V(\frac{\sum_{i=1}^{n}(X_i)}{n}) = \lim_{n\to\infty} \frac{p(1-p)}{n} = 0 \Rightarrow \hat{p}$ *est convergent*

On calcule I(θ)

$$I_n(\theta) = -E(\frac{\delta^2 \ln L}{\delta \theta^2}) = -E(\frac{n}{\theta^2}) + E(2\frac{\sum_1^n X_i}{\theta^3}) \Rightarrow I_n(\theta) = \frac{n}{\theta^2}$$

$$V(\overline{X}) = \frac{\sigma^2}{n} = \frac{1}{I_n(\theta)} \Rightarrow \hat{\theta} \text{ est efficace.}$$

Exercice 5

① *Estimation de p*

On a $\begin{cases} X_i = 1 & \text{avec } \text{Prob}(X_i = 1) = p \\ X_i = 0 \text{ avec} & \text{Prob}(X_i = 0) = 1 - p = q \end{cases}$

La fonction densité de probabilité est donnée par : $f(x, \theta) = p^x(1-p)^{1-x}$

La *vraisemblance* est donnée par

$$L(x_1, \ldots x_n, \theta) = p^{\sum_i^n x_i}(1-p)^{n-\sum_i^n x_i}$$

En passant au logarithme : $\ln L = -\sum_1^n x_i \ln p + (n - \sum_1^n x_i)\ln(1-p)$

En dérivant, on obtient : $$\frac{\delta \ln L}{\delta p} = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p}$$

En annulant la dérivée, on obtient : $\overline{X} = \frac{\sum_{i=1}^n x_i}{n} = \hat{p} = F_n = 0{,}25$

② *Propriétés de* $\hat{p}$

$$E(X) = E(F_n) = p \quad \Rightarrow \hat{p} \text{ est sans biais}$$

$$\lim_{n\to\infty} V(\overline{X}) = \lim_{n\to\infty} V(\frac{\sum_{i=1}^n (X_i)}{n}) = \lim_{n\to\infty} \frac{p(1-p)}{n} = 0 \Rightarrow \hat{p} \text{ est convergent}$$

③ *Distribution de* F_n

$\sum_{i=1}^n X_i \to B(n, p)$; il en est de même pour F_n. Comme $n > 30$ suit une loi normale dont les caractéristiques sont :

$$E(X) = E(F_n) = p \text{ et } V(\overline{X}) = V(\frac{\sum_{i=1}^n (X_i)}{n}) = V(F_n) = \frac{p(1-p)}{n}$$

Chapitre 4

INTERVALLE DE CONFIANCE

☛ *Objectifs*

- ⇨ Introduire la notion d'intervalle de confiance d'un paramètre
- ⇨ estimer par intervalle de confiance une proportion, une différence de proportions
- ⇨ estimer par intervalle de confiance les paramètres de différents types de distribution (moyenne, variance, écart-type ...)
- ⇨ utiliser la loi de Student

☛ *Mots clés*

Intervalle de confiance – intervalle pour p – intervalle pour la différence de proportions – intervalle pour m ; pour σ^2 ; pour σ ; pour une différence de ces paramètres – intervalle de confiance pour un coefficient de corrélation.

I ***Intervalle pour une proportion***
II ***Intervalle pour une moyenne***
III ***Intervalle pour une variance***
IV ***Intervalle pour un coefficient de corrélation***

AUTO-EVALUATION & EXERCICES CORRIGES

Les estimations du chapitre précédent sont des valeurs numériques réelles calculées à partir de valeurs observées. Peut-on avoir certaines informations concernant la précision de ces valeurs ?

La réponse à cette interrogation est due à J. NEYMAN : ***c'est l'estimation par intervalle de confiance ou intervalle de sécurité***. « *On associe au résultat d'un échantillon un intervalle qui doit contenir la vraie valeur inconnue du paramètre, avec une probabilité fixée à l'avance.*»

Soit Θ un estimateur du paramètre inconnu. Il s'agit de trouver deux nombres α et a avec $a > 0$ et $0 < \alpha < 1$ ***tels que*** :

Prob $(\Theta - a \leq \theta \leq \Theta + a) = \text{Prob}\ (t_1 < \theta < t_2) = \alpha$

$[t_1 ; t_2]$ *l'intervalle de confiance*

t_1 *et* t_2 *les limites de confiance*

α : *niveau de probabilité* ou *coefficient de confiance*

$(1 - \alpha)$: *seuil de confiance* ou *niveau de confiance* ou *coefficient de sécurité*.

Pour réaliser une estimation par intervalle de confiance, le problème consiste à déterminer la *loi de distribution* de l'estimateur. Il n'y a pas de règle générale pour le faire. Pour mieux comprendre cette notion, nous étudierons l'exemple de la construction d'un *intervalle de confiance pour une proportion p*, puis pour la *moyenne m et pour l'écart-type* σ d'une *loi de Gauss* et enfin pour le *coefficient de corrélation* ρ d'une *loi bidimensionnelle*.

I - INTERVALLE POUR UNE PROPORTION

Nous savons que le *meilleur estimateur* du paramètre d'une *distribution binomiale* est la *fréquence*. C'est un estimateur *correct*, *convergent*, *efficace* et *asymptotiquement normal*.

I - 1 INTERVALLE POUR p

Nous allons construire un intervalle qui contient la *vraie valeur de p* c'est à dire qui s'en écarte avec une probabilité donnée α (*cette probabilité est un nombre positif, généralement proche de 0 et inférieur à 1*).

✧ *Dans la pratique*, on utilise souvent $\alpha = 0,01$; $\alpha = 0,05$ ou $\alpha = 0,10$. C'est la mesure de la *précision de l'estimation* souhaitée.

1 - 1. 1 Définition

On appelle *intervalle de confiance* pour p, au niveau de probabilité α, un *intervalle centré sur* $p = \frac{k}{n}$ et qui contient p avec une probabilité $1 - \alpha$.

1 - 1.2 Distribution de $\frac{k}{n}$

Cas où n est grand

Comme $\frac{k}{n}$ admet une distribution binomiale $B(n, p)$, cette loi est d'emploi difficile. Supposons la *validité* de son *approximation* par la *loi normale*.

Si **npq = 9** ou encore **np > 15** et **nq > 15** alors

$$\frac{k}{n} \to \mathrm{N}(p, \sqrt{\frac{pq}{n}}) \quad \text{ou} \quad \frac{\frac{k}{n} - p}{\sqrt{\frac{pq}{n}}} \to N(0, 1)$$

On a $\text{Prob}\left(-t_{\alpha/2} \leq \frac{\frac{k}{n} - p}{\sqrt{\frac{pq}{n}}} \leq + t_{\alpha/2}\right) = 1 - \alpha$ où $t_{\alpha/2}$ est une valeur lue dans la table de la *loi normale centrée réduite N(0, 1)*.

En isolant p et en ne tenant pas compte de la *correction de continuité* $(\pm\frac{1}{2})$, on obtient :

$$\text{Prob}\left(p - t_{\alpha/2}\sqrt{\frac{pq}{n}} \leq p \leq p + t_{\alpha/2}\sqrt{\frac{pq}{n}}\right) = 1 - \alpha$$

L'intervalle de confiance au niveau $1 - \alpha$ pour le paramètre p lorsque la *taille* de l'échantillon est *grande* est donné par :

$$\boxed{p \pm \sqrt{\frac{pq}{n}}}$$

Cet intervalle dépend de p, comment le déterminer ?

⌘ *Quatre méthodes généralement utilisées*

➲ *Méthode de l'ellipse*

$$-t_{\alpha/2}\sqrt{\frac{pq}{n}} < \frac{k}{n} - p < t_{\alpha/2}\sqrt{\frac{pq}{n}} \quad \Leftrightarrow \quad (\frac{k}{n} - p)^2 < t^2 \frac{pq}{n}$$

Pour **n** et **t** fixés, cette inégalité est vérifiée pour tous les points situés à l'intérieur d'une *ellipse* et seules sont à *conserver* les parties correspondantes à :

$$0 \leq p \leq 1 \text{ et } 0 \leq \frac{k}{n} \leq 1$$

A partir de la proportion d'ellipse ainsi délimitée, on détermine pour $\frac{k}{n}$ les limites de l'intervalle de confiance.

✧ L'ensemble des ellipses ainsi obtenues, pour un *seuil* de confiance *donné*, en faisant *varier n*, forme un ***abaque***.

✧ En *pratique*, on dispose *d'abaques*, pour des *valeurs de* α et *des valeurs de n*. On *lit directement* sur l'abaque l'intervalle correspondant au $\frac{k}{n}$ *observé*.

➲ *Approximation de p :*

En remplaçant *p* par *son estimateur* sans biais et efficace $\frac{k}{n}$ et avec $q = 1 - p$, on a : $pq = \frac{k}{n}(1 - \frac{k}{n})$ d'où l'intervalle :

$$[\frac{k}{n} - t_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}} ; \ \frac{k}{n} + t_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}} \ ]$$

Exemple

Pour fixer les idées, on veut estimer la proportion de conducteurs d'automobiles qui portent la ceinture de sécurité. Sur un échantillon de 160 conducteurs observés à l'entrée d'une agglomération, on a noté qu'il y en a 96 qui la portent. Construire un intervalle de confiance au niveau 95 % pour estimer la vraie proportion des conducteurs portant la ceinture.

◆ Pour *un niveau 95 % et n = 100*, np > 15 et nq >15 on a :

$$\text{Prob}\left| \frac{k}{n} - 1{,}96\sqrt{\frac{1}{n}\frac{k}{n}\left(1-\frac{k}{n}\right)} \le p \le \frac{k}{n} + 1{,}96\sqrt{\frac{1}{n}\frac{k}{n}\left(1-\frac{k}{n}\right)} \right| = 0{,}95$$

d'où $p \in [0{,}5\ ;\ 0{,}7]$. Cet intervalle contient la *vraie valeur de p* avec une *probabilité* de *0,95*.

➲ *Intervalle approché par excès*

On sait qu'un *produit ab* de deux *nombres positifs* dont la *somme* est *constante* est *maximum* lorsque les *deux nombres sont égaux*.

C'est le cas du produit pq d'où $p = q = \frac{1}{2}$. En remplaçant pq par sa *valeur maximum* $\frac{1}{4}$, on obtient l'intervalle : $\frac{k}{n} \pm t_{\alpha/2}\sqrt{\frac{1}{4n}}$

➲ *Autre méthode : loi arc sinus*

On démontre que la variable aléatoire $2arc\ sin\frac{k}{n}$ est *asymptotiquement normale* de moyenne $m = 2arc\ sin\sqrt{p}$ et d'écart-type $\sigma = \frac{1}{\sqrt{n}}$

On a alors :

$$\text{Prob}\left(-t_{\alpha/2} \le \frac{2\arcsin\sqrt{\frac{k}{n}} - 2\arcsin\sqrt{p}}{\frac{1}{\sqrt{n}}} \le t_{\alpha/2}\right) = 1-\alpha$$

L'intervalle de confiance est donné par $[2arc\ sin\frac{k}{n} \pm t_{\alpha/2}\frac{1}{\sqrt{n}}]$.

Des *tables* permettent de passer *de x à 2arc sin* $\sqrt{x}$ et *réciproquement*.

Remarques

☞ La méthode de l'ellipse est celle qui est valable théoriquement pour la détermination des intervalles de confiance.

☞ Les abaques permettent de déterminer $\frac{k}{n}$ connaissant p

☞ Pour un *risque* consenti α et une *précision e* désirée, la *taille n* de l'échantillon est donnée par : $n = \frac{t_{\alpha/2}^2 pq}{e^2}$

Exemple

Soit $n = 62$ et $k = 12$ avec $\alpha = 0,05$

Sur la ligne $k = 62$ et dans la colonne $n - k = 62 - 12 = 50$, on lit, au niveau de *probabilité 0,05*, les *limites* de l'intervalle de confiance *0,104* et *0,314*.

I - 2 INTERVALLE POUR UNE DIFFERENCE

I - 2. 1 Définition

On appelle intervalle de confiance, au niveau de *probabilité* α, l'intervalle centré sur $\frac{k_1}{n_1} - \frac{k_2}{n_2}$ qui *contient la vraie* valeur de la *différence* avec une *probabilité de 1- α.*

I - 2. 2 Propriétés

La distribution de $\frac{k_1}{n_1} - \frac{k_2}{n_2}$ est *approximativement normale*

$$\frac{k_1}{n_1} - \frac{k_2}{n_2} \rightarrow N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

d'où $$\text{Prob}\left[\left(\frac{k_1}{n_1} - \frac{k_2}{n_2}\right) - (p_1 - p_2) \leq t_{\alpha/2} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \right] = 1 - \alpha$$

$p_1$et p_2 étant *inconnues*, il faut alors *les estimer par leur fréquence respective* pour obtenir les limites de l'intervalle de confiance.

Les limites de l'intervalle de confiance pour $(p_1 - p_2)$ sont :

$$\left(\frac{k_1}{n_1} - \frac{k_2}{n_2}\right) \pm t_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

avec $\hat{p}_1 = \frac{k_1}{n_1}$; $\hat{p}_2 = \frac{k_2}{n_2}$; $\hat{q}_1 = 1 - \hat{p}_1$; $\hat{q}_2 = 1 - \hat{p}_2$

Exemple

Une enquête effectuée dans une administration indique que 320 hommes mariés sur 450 choisis au hasard et 120 célibataires choisis au

hasard sur 200 souhaitent travailler en séance unique. Donner un intervalle de confiance pour la différence, au coefficient de sécurité $\alpha = 0,95$.

$$(0,71-0,6) \pm 1,96\sqrt{\frac{0,71\times0,29}{450}+\frac{0,6\times0,4}{200}}$$ d'où l'intervalle [0,03 ; 0,19]

Remarque

☞ ***Echantillons appariés***

Des échantillons sont dits appariés lorsqu'il y a une *relation biunivoque entre eux*. Il s'agit par exemple de la comparaison de deux groupes constitués par les mêmes sujets ou encore lorsque les échantillons présentent une ou plusieurs caractéristiques communes (même répartition selon l'âge, le poids, la taille, le sexe,...). Dans ce cas, on considère la *différence des valeurs observées* et *on procède* comme pour *une proportion*.

II - INTERVALLE POUR UNE MOYENNE

✓ Soit une population ***P***, de moyenne *m* et d'écart-type σ.

✓ On prélève au hasard, avec remplacement et de manière indépendante, un échantillon aléatoire d'effectifs *n* grand.

✓ A toute valeur typique de l'échantillon, on peut associer une variable aléatoire. La loi de probabilité de cette variable est appelée *distribution échantillée* de cette valeur typique

✓ D'après le théorème central-limite cette distribution est asymptotiquement normale.

II - 1 INTERVALLE POUR m

Selon la nature de la distribution, on se propose de déterminer dans chaque cas, l'intervalle de confiance.

II - 1. 1 DISTRIBUTION QUELCONQUE

Pour déterminer un intervalle de confiance pour la moyenne *m* de la population *P deux cas* sont à distinguer selon que la *variance est connue* ou *non*.

a) **σ connu**

❶ ***Analyse***

On a approximativement

$$\text{Prob}\left(-t_{\alpha/2} \leq \frac{\bar{x} - m}{\frac{\sigma}{\sqrt{n}}} \leq t_{\alpha/2}\right) = 1-\alpha$$

$t_{\alpha/2}$ est la valeur de la variable centrée réduite $T \rightarrow N(0,1)$ qui correspond au niveau α.

❷ ***Intervalle***

L'intervalle qui contient la vraie valeur de m avec la probabilité $1-\alpha$ est déterminé par :

$$\boxed{\bar{x} \pm t_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

❸ **Exemple**

Supposons que l'on extrait un échantillon ayant les caractéristiques suivantes : La taille n = 25 ; la moyenne $\bar{x} = 3{,}2$ et l'écart-type $\sigma = 2$.
Déterminer un intervalle de confiance au niveau α = 95 %.

On sait que les bornes sont données par :

$$\boxed{\bar{x} \pm t_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

En prenant $t \# 2$ et en remplaçant σ et n par leur valeur respective dans l'expression précédente, on obtient : $m \in [2{,}4 ; 4]$ ce qui signifie qu'il y a 95 chances sur 100 pour que la moyenne m se trouve dans cet *intervalle*.

b) **σ inconnu**

❶ ***Analyse***

Pour n grand, l'écart-type s de l'échantillon peut être considéré comme une bonne approximation de σ. L'approximation de σ conduit à la distribution approchée de

$\bar{x} \rightarrow N\left(m, \frac{s}{\sqrt{n}}\right)$ et en *remplaçant σ par s*, on obtient la relation :

$$\text{Prob}\left(-t_{\alpha/2} \leq \frac{\bar{x} - m}{\frac{s}{\sqrt{n}}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

❷ ***Intervalle***

L'intervalle pour m, au niveau α, a pour bornes

$$\boxed{\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}}$$

II - 1. 2 DISTRIBUTION NORMALE

✓ P admet dans ce cas une distribution normale $N(m, \sigma)$.

✓ Comme dans le cas précédent, on prélève un échantillon aléatoire d'effectif n, de moyenne $\bar{x}$ et d'écart-type s.

Pour déterminer l'intervalle de confiance deux cas à examiner :

a) **σ connu**

❶ ***Analyse***

✧ Aucune approximation n'est nécessaire dans ce cas

✧ $\bar{x}$ suit la distribution $N(m, \frac{\sigma}{\sqrt{n}})$

$$\text{Prob}\left(-t_{\alpha/2} \leq \frac{\bar{x} - m}{\frac{\sigma}{\sqrt{n}}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

$t_{\alpha/2}$ est la valeur de la variable centrée réduite $T \rightarrow N(0,1)$ qui correspond au niveau α.

❷ ***Intervalle***

L'intervalle de confiance est le même que celui déterminé pour une distribution quelconque :

$$\boxed{\bar{x} \pm t_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

b) **σ inconnu**

❶ ***Analyse***

✧ Il n'est plus possible de remplacer l'écart-type par son estimation S, cette approximation n'étant valable que si la taille de l'échantillon est grande ($n > 30$), *ce qui n'est pas nécessairement le cas.*

✧ On utilise alors la *distribution exacte* :

$$\frac{\bar{x}-m}{\frac{S}{\sqrt{n}}} = \frac{\bar{x}-m}{\frac{s}{\sqrt{n-1}}} \rightarrow t_{n-1}$$

C'est la variable de ***Student à n – 1 degrés de liberté***

❷ ***Intervalle***

L'intervalle pour m, au niveau α, a pour bornes

$$\boxed{\bar{x} \pm t_{n-1;\alpha/2} \frac{S}{\sqrt{n}}}$$

II - 2 INTERVALLE POUR UNE DIFFERENCE

✓ Soit P_1 et P_2 deux populations quelconques de moyennes respectives m_1 et m_2 et d'écarts types respectifs σ_1 et σ_2.

✓ On prélève deux échantillons aléatoires d'effectifs $n_1 \geq 30$ et $n_2 \geq 30$ ayant respectivement $\bar{x}_1$ et $\bar{x}_2$ comme moyennes arithmétiques s^2_1 et s^2_2 comme variances.

✓ Pour déterminer l'intervalle de confiance ***plusieurs cas à examiner*** selon la *nature de la distribution* :

II - 2. 1 DISTRIBUTION QUELCONQUE

Pour déterminer l'intervalle de confiance ***deux cas à étudier*** :

a) σ_1 **et** σ_2 **connus**

❶ ***Analyse***

✧ On sait que

$$\bar{x}_1 \rightarrow N(m_1, \frac{\sigma_1}{\sqrt{n_1}}) \quad \text{et} \quad \bar{x}_2 \rightarrow N(m_2, \frac{\sigma_2}{\sqrt{n_2}})$$

✧ La variable

$$\bar{x}_1 - \bar{x}_2 \rightarrow N(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

✧ On définit la variable *normale centrée réduite*

$$\frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}} \rightarrow N(0,1)$$

$$\text{Pob}\left(\frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}} \right| \leq t_{\alpha/2}) = 1 - \alpha$$

❷ ***Intervalle***

De la relation précédente, on en déduit l'intervalle de confiance pour $(m_1 - m_2)$ au niveau α :

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}})$$

b) σ_1 **et** σ_2 **inconnus**

❶ ***Analyse***

✧ On remplace les écarts-types des formules précédentes par les variances des échantillons respectifs S^2_1 et S^2_2

❷ ***Intervalle***

L'intervalle de confiance pour $(m_1 - m_2)$ au niveau α :

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \sqrt{\frac{S^2_1}{n_1} + \frac{S^2_2}{n_2}})$$

II - 2. 2 DISTRIBUTION NORMALE

✓ Les deux *populations* P_1 et P_2 *étant normales* de moyennes respectives m_1 et m_2 et d'écart-type $\sigma_1 = \sigma_2 = \sigma$. Comme dans le cas précédent, on prélève deux échantillons aléatoires d'effectifs $n_1 \geq 30$ et $n_2 \geq 30$ ayant respectivement $\bar{x}_1$ et $\bar{x}_2$ comme moyennes arithmétiques s^2_1 et s^2_2 comme variances. *Pour déterminer l'intervalle de confiance, deux cas à examiner* :

a) σ_1 **et** σ_2 **connus**

❶ ***Analyse***

On procède comme dans le cas de la distribution quelconque et on remplace $\sigma_1 = \sigma_2 = \sigma$

On sait que :

$$\bar{x}_1 \to N(m_1, \frac{\sigma}{\sqrt{n_1}}) \quad \text{et} \quad \bar{x}_2 \to N(m_2, \frac{\sigma}{\sqrt{n_2}})$$

$$\text{et} \qquad \frac{\bar{x}_1 - \bar{x}_2 - (m_1 - m_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \to N(0,1)$$

❷ ***Intervalle***

L'intervalle de confiance pour $(m_1 - m_2)$, au niveau α, *a pour bornes* :

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \ \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}})$$

a) σ_1 **et** σ_2 **inconnus**

❶ ***Analyse***

✧ Comme dans le cas précédent $\sigma_1 = \sigma_2 = \sigma$

$$\frac{\bar{x}_1 - \bar{x}_2 - (m_1 - m_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \rightarrow N(0,1)$$

mais cette variable centrée réduite fait *intervenir* σ *inconnu*, on procède alors comme suit : En utilisant les *propriétés connues* suivantes

- $\frac{n_1 s^2_1}{\sigma^2} \rightarrow \chi^2_{n_1-1}$ et $\frac{n_2 s^2_2}{\sigma^2} \rightarrow \chi^2_{n_2-1}$
- $\chi^2_{n_1-1} + \chi^2_{n_2-1} = \chi^2_{(n_1-1)+(n_2-1)}$
- $\frac{n_1 s^2_1 + n_2 s^2_2}{\sigma^2} \rightarrow \chi^2_{n_1+n_2-2}$

et la variable de *Student*, on obtient alors :

$$t_{n_1+n_2-2} \rightarrow \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n_1+n_2-2}}{n_1+n_2-2}}} = \frac{(\bar{x}_1-\bar{x}_2) - (m_1-m_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

avec $S^2 = \frac{n_1 s^2_1 + n_2 s^2_2}{n_1+n_2-2}$ qui est un estimateur sans biais de σ^2

❷ ***Intervalle***

L'intervalle de confiance pour $(m_1 - m_2)$, au *niveau* α, *a pour bornes* :

$$\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2;\,\alpha/2} \ S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Remarque

☞ ***Echantillons appariés*** :

Pour comparer les moyennes de *deux séries appariées* de *faibles effectifs*, on forme pour *chaque paire la différence entre les moyennes* puis on procède comme pour *une seule moyenne*.

III - INTERVALLE POUR UNE VARIANCE

✓ Soit une population P admettant une distribution, de moyenne m et d'écart-type σ.

✓ On utilise les estimateurs sans biais de la variance

✓ On considère un échantillon de grande taille ($n \geq 50$)

III - 1 INTERVALLE POUR σ^2

Plusieurs cas à examiner, selon la nature de la distribution de la population.

III - 1. 1 DISTRIBUTION QUELCONQUE

❶ ***Analyse***

✧ Pour *n suffisamment grand* $n \geq 50$

$$S^2 \to N(\sigma^2, \sqrt{\frac{\mu_4 - \sigma^4}{n}}) \quad \text{et} \quad S \to N(\sigma, \sqrt{\frac{\mu_4 - \mu^2_2}{4n\mu_2}})$$

μ_2; σ^2 et μ_4 doivent être estimés par $\hat{\mu}_2 = S^2$ et $\hat{\mu}_4 = m_4$.

La variable $$\frac{S^2 - \sigma^2}{\sqrt{\frac{\hat{\mu}_4 - \hat{\mu}^2_2}{n}}}$$

peut être assimilée à une variable centrée réduite.

Soit :

$$\text{Prob}\left(\left| \frac{S^2 - \sigma^2}{\sqrt{\frac{\hat{\mu}_4 - \hat{\mu}^2_2}{n}}} \right| \leq t_{\alpha/2} \right) = 1 - \alpha$$

on obtient :

$$\text{Prob}\left(S^2 - t_{\alpha/2}\sqrt{\frac{m_4 - S^4}{n}} \leq \sigma^2 \leq S^2 + t_{\alpha/2}\sqrt{\frac{m_4 - S^4}{n}} \right) \leq 1 - \alpha$$

❷ ***Intervalle***

✧ *L'intervalle* de *confiance pour* σ^2, au *niveau* α, *a pour bornes* :

$$S^2 \pm t_{\alpha/2}\sqrt{\frac{m_4 - S^4}{n}}$$

✧ *L'intervalle* de *confiance pour* σ au *niveau* α peut s'obtenir en prenant les racines carrées des limites de l'intervalle précédent ou encore à partir de :

$$S \to N(\sigma, \sqrt{\frac{\mu_4 - \mu^2_2}{4n\mu_2}}),$$

d'où l'intervalle : $S \pm t_{\alpha/2}\sqrt{\frac{m_4 - S^4}{4nS^2}}$

Exemple

Construire un intervalle de confiance à 95% pour σ^2 et σ dans le cas suivant :

n = 120 ; s^2 = 17, le moment d'ordre 4 est égal à 322.

◆ En appliquant les formules précédentes, on obtient les intervalles respectifs :

$\sigma^2 \in$ [15,97 ; 18,03] et $\sigma \in$ [3,998 ; 4,248]

III - 1. 2 DISTRIBUTION NORMALE

Rappelons les résultats suivants qui vont permettre de trouver les différents intervalles de confiance :

✓ $\frac{ns^2}{\sigma^2} \to \chi^2_{n-1}$

✓ $\frac{S^2_1}{S^2_2} = \frac{\frac{n_1 s^2_1}{n_1 - 1}}{\frac{n_2 s^2_2}{n_2 - 1}} = F_{n_1-1;n_2-1}$ à condition que $\sigma^2_1 = \sigma^2_2$

loi de Fisher-Snedecor à $(n_1 - 1)$ *et* $(n_2 - 1)$ *degrés de liberté*

✓ $n \geq 30$ $\sqrt{2\chi^2_{n-1}} \to N(\sqrt{2n-3}; 1)$

❶ ***Analyse***

✧ On détermine les deux valeurs $\chi^2_{n-1,\alpha/2}$ et $\chi^2_{n-1,1-\alpha/2}$ telle que :

$$\text{Prob}(\chi^2_{n-1,\alpha/2} \leq \frac{ns^2}{\sigma^2} \leq \chi^2_{n-1;1-\alpha/2}) = 1-\alpha$$

On en déduit :

$$\text{Prob}\left(\frac{ns^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{ns^2}{\chi^2_{n-1;\alpha/2}} \right) = 1-\alpha$$

❷ ***Intervalle***

✧ *L'intervalle* de *confiance pour* σ^2, au *niveau* α, *a pour bornes* :

$$\left[\frac{ns^2}{\chi^2_{n-1;1-\alpha/2}} \; ; \; \frac{ns^2}{\chi^2_{n-1;\alpha/2}} \right]$$

✧ *L'intervalle* de *confiance pour* σ, au *niveau* α, peut s'obtenir en prenant les racines carrées des limites de l'intervalle précédent.

$$\left[\frac{s\sqrt{n}}{\sqrt{\chi^2_{n-1;1-\alpha/2}}} \; ; \; \frac{s\sqrt{n}}{\sqrt{\chi^2_{n-1;\alpha/2}}} \right]$$

Remarques

☞ Pour les grands échantillons, on utilise l'approximation :

$$n \geq 30 \qquad \sqrt{2\chi^2_{n-1}} - \sqrt{2n-3} \rightarrow N(0,\ 1)$$

☞ *L'intervalle* de *confiance pour* σ, au *niveau* α, est donné par :

$$\left[\frac{s\sqrt{2n}}{\sqrt{2n-3} + t_{\alpha/2}} \; ; \; \frac{s\sqrt{2n}}{\sqrt{2n-3} - t_{\alpha/2}} \right]$$

III – 1. 3 DISTRIBUTION APPROXIMATIVEMENT NORMALE

✓ Sous certaines conditions, les moments à l'origine, les moments centrés, ... sont des *variables asymptotiquement normales*.

✓ Pour une variable normale ou asymptotiquement normale la distribution est déterminée par la moyenne et l'écart-type.

❶ ***Analyse***

Si la population est quelconque mais que la taille de l'échantillon est suffisamment grande, on sait que :

$$S^2 \to N(\sigma^2, \sqrt{\frac{\mu_4 - \sigma^4}{n}})$$

En outre, si la population est normale le *coefficient* de *Pearson* vaut 3 et on a :

$$\mu_4 = 3\sigma^4 = 3\mu_2^2$$

$$\mu_4 - \mu_2^2 = 3\mu_2^2 - \mu_2^2 = 2\sigma^4$$

Même si la population n'est pas normale et si le *coefficient d' aplatissement* de *Pearson* vaut 3 alors elle est *approximativement normale*.

❷ ***Intervalle***

On peut remplacer la variance S^2 par $\frac{2\sigma^4}{n}$ d'où

$$S^2 \to N(\sigma^2, S^2\sqrt{\frac{2}{n}}) \text{ et } S \to N(\sigma, S\sqrt{\frac{1}{2n}})$$

Pour déterminer l'intervalle de confiance, *au niveau* α, on utilise la distribution approchée de S^2 :

$$\text{Prob}\left(\frac{S^2 - \sigma^2}{S^2\sqrt{\frac{2}{n}}} \right| \leq t_{\alpha/2}) = 1 - \alpha$$

et les bornes de l'intervalle sont respectivement pour σ^2 et pour σ

$$S^2 \pm t_{\alpha/2}S^2\sqrt{\frac{2}{n}} \qquad \text{et} \qquad S \pm t_{\alpha/2}S\sqrt{\frac{1}{2n}}$$

Remarques

☞ Si la population est normale, on utilise la distribution exacte

☞ Le raisonnement effectué pour S^2 demeure valable pour S

☞ On estime la variance par celle de l'échantillon.

III - 2 INTERVALLE POUR UNE DIFFERENCE

✓ Soit P_1 et P_2 deux populations quelconques de moyennes respectives m_1 et m_2 et d'écarts types respectifs σ_1 et σ_2

✓ On prélève deux échantillons aléatoires d'effectifs $n_1 \geq 30$ et $n_2 \geq 30$ ayant respectivement $\bar{x}_1$ et $\bar{x}_2$ comme moyennes arithmétiques s^2_1 et s^2_2 comme variances.

✓ L'intervalle pour la différence des variances dépend de la nature de la distribution d'où *plusieurs cas à étudier*:

III - 2. 1 DISTRIBUTION QUELCONQUE

❶ ***Analyse***

$$S^2_1 \to N(\sigma_1^2, \sqrt{\frac{\mu^{(1)}_4 - \sigma_1^4}{n_1}}) \qquad \text{et} \qquad S^2_2 \to N(\sigma_2^2, \sqrt{\frac{\mu^{(2)}_4 - \sigma_2^4}{n_2}})$$

Il en est de même pour les écarts types

$$S_1 \to N(\sigma_1, \sqrt{\frac{\mu^{(1)}_4 - \sigma_1^4}{4n_1\sigma^2_1}}) \qquad \text{et} \qquad S_2 \to N(\sigma_2, \sqrt{\frac{\mu^{(2)}_4 - \sigma_2^4}{4n_2\sigma^2_2}})$$

(1) et (2) les échantillons respectifs

μ_2, σ^2 et μ_4 doivent être remplacés par leurs estimations respectives.

❷ ***Intervalle***

Les *intervalles de confiance, au niveau* α, s'obtiennent par :

$$S^2_1 - S^2_2 \pm t_{\alpha/2}\sqrt{\frac{\hat{\mu}^{(1)}_4 - \hat{\sigma}_1^4}{n_1} + \frac{\hat{\mu}^{(2)}_4 - \hat{\sigma}_2^4}{n_2}} \quad \textit{et} \quad S_1 - S_2 \pm t_{\alpha/2}\sqrt{\frac{\hat{\mu}^{(1)}_4 - \hat{\sigma}_1^4}{4n_1\hat{\sigma}^2_1} + \frac{\hat{\mu}^{(2)}_4 - \hat{\sigma}_2^4}{4n_2\hat{\sigma}_2}}$$

III - 2. 2 DISTRIBUTION APPROXIMATIVEMENT NORMALE

✓ Les populations sont approximativement normales et les échantillons aléatoires

❶ ***Analyse***

On a :

$$S^2_1 \to N(\sigma_1^2, \sqrt{\frac{2\sigma_1^4}{n_1}}) \; ; \; S^2_2 \to N(\sigma_2^2, \sqrt{\frac{2\sigma_2^4}{n_2}})$$

$$S_1 \to N(\sigma_1, \sqrt{\frac{\sigma_1^2}{2n_1}}) \; ; \qquad S_2 \to N(\sigma_2, \sqrt{\frac{\sigma_2^2}{2n_2}})$$

❷ ***Intervalle***

Les intervalles de confiance, *au niveau* α, s'obtiennent par remplacement des écarts types par leurs estimations non biaisées :

$$S^2_1 - S^2_2 \pm t_{\alpha/2}\sqrt{\frac{2S_1^4}{n_1}+\frac{2S_2^4}{n_2}} \quad \text{et} \quad S_1 - S_2 \pm t_{\alpha/2}\sqrt{\frac{S_1^2}{2n_1}+\frac{S_2^2}{2n_2}}$$

Exemple

Construire un intervalle de confiance à 95% pour une différence de variance de deux populations presque normales dans le cas où

$n_1 = 180$; $n_2 = 160$; $s^2_1 = 47$; $s^2_2 = 36$

◆ En appliquant les formules précédentes, on obtient les intervalles respectifs :

$\sigma^2_1 \in [37{,}298 ; 57{,}702]$ et $\sigma^2_2 \in [3{,}998 ; 4{,}248]$

d'où $\sigma^2_1 - \sigma^2_2 \in [9{,}187 ; 12{,}813]$

IV - COEFFICIENT DE CORRELATION

Dans le manuel « ***Statistique & Probabilités*** », nous avons étudié les distributions à deux dimensions. Soit une population normale à *deux dimensions* (X,Y) dont les *paramètres* sont respectivement m_1 ; m_2 ; σ_1 ; σ_2 ; ρ.

A partir d'un échantillon n, on se propose d'estimer *le coefficient de corrélation* ρ. On utilise le coefficient de *corrélation de l'échantillon* r.

❶ *Analyse*

✧ La distribution de r n'est pas utilisable dans la pratique. On montre que :

* r suit *approximativement la loi normale*

$$N(\rho, \frac{1-\rho^2}{\sqrt{n-1}})$$

* r est un estimateur convergent de ρ.

✧ On introduit la *transformation de Fisher* :

$$Z' = \frac{1}{2}\log_e \frac{1+r}{1-r}$$

✧ Pour n *suffisamment grand* ($n \geq 25$), Fisher a démontré que Z' admet une *distribution approximativement normale* :

$$E(Z') = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \quad \sigma(Z') = \frac{1}{\sqrt{n-3}} \quad \text{et que} \quad Z' \rightarrow N[E(Z'); \frac{1}{\sqrt{n-3}}]$$

✧ Pour une *population normale* $\rho = 0$, on montre que la variable aléatoire

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \rightarrow t_{n-2}$$ suit une *loi de Student à (n – 2) degrés de liberté*.

❷ ***Intervalle***

Les limites de l'intervalle de confiance pour ρ, au niveau de probabilité α, sont :

$$r \pm t_{\alpha/2} \frac{1 - r^2}{\sqrt{n-1}}$$

✧ On construit un *intervalle de confiance autour de E(Z')* donné par :

$$Z' - t_{\alpha/2} \frac{1}{\sqrt{n-3}} \leq E(Z') \leq Z' + t_{\alpha/2} \frac{1}{\sqrt{n-3}}$$

Z' est la *transformée de r* (table de Fisher) et $t_{\alpha/2}$ est donnée par la loi normale *centrée réduite*.

✧ Des limites précédentes, on obtient, en les *transformant en r, celles de* ρ, en utilisant la *table de Fisher en sens inverse* (*Table III*).

Auto-évaluation ?

1- Les limites de l'intervalle de confiance sont des valeurs fixes

a) Vrai b) Faux

2- Le niveau de confiance est toujours associé à un paramètre θ de la population à estimer

a) Vrai b) Faux

3- Un intervalle de confiance est toujours centré sur la valeur de l'estimateur

a) Vrai b) Faux

4- Plus le niveau de confiance associé à l'intervalle est élevé plus l'amplitude de ce dernier est élevée

a) Vrai b) Faux

5- La moyenne μ appartient à un intervalle de confiance $[a\ ;\ b]$. Parmi les réponses suivantes trouver la bonne interprétation :

1) Il y a 95 chances sur 100 pour que la vraie valeur de la moyenne soit dans l'intervalle $a \leq \mu \leq b$.

2) On est sûr à 95 % que la vraie valeur de la moyenne est comprise entre a et b.

3) L'intervalle de confiance $a \leq \mu \leq b$ a 95 chances sur 100 de contenir la vraie valeur de μ

6- Construire un intervalle de confiance à 95 % pour σ^2 *et* σ :

1) Population normale

$n = 15$ $s^2 = 7{,}5$

2) Population quelconque

$n = 110$ $s^2 = 18$ $m_4 = 358$

3) Population approximativement normale

$n = 110$ $s^2 = 18$ $m_4 = 358$

7- Soit P une population d'individus tel que chacun possède ou non une certaine caractéristique avec les probabilités respectives p et $q = 1 - p$. Construire un intervalle de confiance pour p

8- Soient P_1 et P_2 deux populations caractérisées respectivement par les variables aléatoires indépendantes de moyennes m_1 pour X et m_2 pour Y. Soient un échantillon de taille n_1 issu de P_1 et un échantillon de taille n_2 issu de P_2. Notons $\bar{x}$; s^2_1 et $\bar{y}$; s^2_2 les moyennes et les variances échantillons de ces données.

Construire un intervalle de confiance pour $(m_1 - m_2)$.

Exercices ?

Exercice 1

La distribution de X « nombre quotidien de paquets de cigarettes par fumeur » sur un échantillon de 100 fumeurs est donnée par le tableau suivant :

X	0,5	1	1,5	2	3
n_i	20	48	15	12	5

1- Calculer l'espérance mathématique et la variance de l'échantillon.
2- Donner un intervalle de confiance de l'espérance de X centré sur la moyenne de l'échantillon, au taux de confiance de 95 %.
3- Déterminer la taille de l'échantillon nécessaire pour connaître la moyenne au risque de 5 % avec une erreur absolue inférieure à 1 %.

Exercice 2

Le taux de glycémie chez un sujet sain est assimilé à une variable aléatoire normale X. L'examen d'un échantillon aléatoire de 16 sujets sains a donné la distribution suivante :

Taux de glycémie	0,85	0,88	0,90	0,92	0,95	1
Nombre de sujets	1	2	4	3	3	3

1- Calculer l'espérance mathématique et la variance de l'échantillon puis donner une estimation de l'espérance mathématique de X et de sa variance.
2- Donner l'intervalle de confiance de l'espérance de X centré sur la moyenne de l'échantillon, au risque $\alpha = 0,02$.

Exercice 3

① Un premier échantillon de taille $n_1 = 9$ issu d'une loi normale $N(\mu_1, \sigma_1)$ a donné $\bar{x}_1 = 42$ *et* $s^2_1 = 38$

Déterminer un intervalle de confiance pour σ_1^2 au niveau de 90%.

② Un deuxième échantillon de taille $n_2 = 6$ issu d'une loi normale $N(\mu_2, \sigma_2)$ a donné $\bar{x}_2 = 36$ *et* $s^2_2 = 42$

Déterminer un intervalle de confiance pour σ_2^2 au niveau de 90%.

③ Comparer les résultats trouvés en ① et ②.

④ Trouver un intervalle de confiance pour $\frac{\sigma_1^2}{\sigma_2^2}$ au niveau 90 %.

⑤ Peut-on trouver un intervalle de confiance pour $\mu_1 - \mu_2$? Donner un intervalle de confiance pour $\mu_1 - \mu_2$ au niveau 90 %.

D'après série 1996-1997, ISG Tunis.

Exercice 4

On considère deux populations presque normales $N(\mu_1, \sigma_1)$ et $N(\mu_2, \sigma_2)$ dans lesquelles on prélève deux échantillons, ayant les caractéristiques suivantes :

$n_1 = 125$ $\quad s^2_1 = 49$ $\quad n_2 = 150$ $\quad s^2_2 = 36$

Construire un intervalle de confiance à 95 % pour $(\sigma^2_1 - \sigma^2_2)$ et $(\sigma_1 - \sigma_2)$.

Exercice 5

Construire un intervalle de confiance pour $\frac{\sigma_1^2}{\sigma_2^2}$, au niveau 90 %

$n_1 = n_2 = 16$ $\quad s^2_1 = 50$ $\quad s^2_2 = 16$

Exercice 6

Dans la salle d'attente d'un service d'un hôpital A, un échantillon de 29 malades montre que le temps moyen d'attente est de 42 minutes avec un écart-type de 9 minutes. Dans un autre hôpital B, un échantillon de 20 malades montre que le temps moyen d'attente est de 35 minutes avec un écart-type de 7 minutes. On suppose que la distribution du temps d'attente dans les deux hôpitaux suit une loi normale.

1- Donner un intervalle de confiance à 90 % pour le temps d'attente moyen dans le service de l'hôpital A.
2- Donner un intervalle de confiance à 90 % pour le temps d'attente moyen dans le service de l'hôpital B.
3- Donner un intervalle de confiance à 90 % pour la différence entre les temps d'attente moyens dans les services des hôpitaux, en supposant que les écarts-types des temps d'attente sont inconnus mais égaux.

D'après examen juin 1997, ISG Tunis.

Exercice 7

Dans une enquête portant sur 100 étudiants en biologie, 60 souhaitent s'inscrire au DEA.

1- Trouver un intervalle de confiance, au seuil de 95 %, de la proportion des étudiants qui souhaitent s'inscrire au DEA.
2- Combien d'étudiants doit-on interroger pour que l'intervalle de confiance soit $\pm$ 1 %, au seuil de confiance 95 % ?
3- Avec les étudiants interrogés, quel est le seuil de confiance qui donne un intervalle de $\pm$ 1 %.

Exercice 8

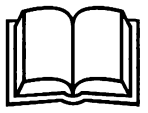
On veut estimer la moyenne d'une variable aléatoire X suivant une loi normale $N(\mu, \sigma)$ de variance $\sigma^2 = 6,25$ à l'aide d'un échantillon aléatoire de taille $n = 100$. La moyenne de l'échantillon $\bar{x} = 4,3$

1- Construire un intervalle de confiance d'un niveau égal à 0,95. Quelle est la signification de cette méthode ?
2- Comment construire un tel intervalle, si l'on ne connaît pas la variance de X, mais seulement la variance empirique de l'échantillon $s^2 = 6,76$.

Exercice 9

Pour un échantillon normal d'effectif $n = 50$, on trouve un coefficient de corrélation $r = 0,3$.

Déterminer un intervalle de confiance pour le coefficient de corrélation ρ de la population, en prenant $\alpha = 0,05$.



Corrigés

Auto-évaluation

1- b

2- b

3- b

4- b

5- 3

6-

① *Population normale*

On sait que : $\frac{ns^2}{\sigma^2} \rightarrow \chi^2_{n-1}$

$$\text{Prob}\left(a < \frac{112,5}{\sigma^2} < b\right) = 0,95$$

soit $\text{Prob}\left(\frac{112,5}{\sigma^2} < a\right) = 0,025$ et $\text{Prob}\left(\frac{112,5}{\sigma^2} < b\right) = 0,975$

La table du khi-deux nous donne $a = 5,01$ et $b = 24,74$

D'où $\text{Prob}\left(5,01 < \frac{112,5}{\sigma^2} < 24,74\right) = 0,95$

$\Rightarrow$ $\text{Prob}(4,547 < \sigma^2 < 22,455) = 0,95$

Les intervalles de confiance cherchés sont donc :

$$\sigma^2 \in [4,547 ; 22,455] \text{ et } \sigma \in [2,132 ; 4,739]$$

② *Population quelconque*

On sait que: $s^2 \rightarrow N\left(\sigma^2, \sqrt{\frac{m_4 - s^4}{n}}\right)$ or $T = \frac{18-\sigma^2}{0,556} \rightarrow N(0,1)$

$$\text{Prob}\left(-1,96 < \frac{18-\sigma^2}{0,556} < 1,96\right) = 0,95 \Rightarrow \text{Prob}(16,91 < \sigma^2 < 19,09) = 0,95$$

$s \rightarrow N\left(\sigma, \frac{1}{2s}\sqrt{\frac{m_4 - s^4}{n}}\right)$ soit $\text{Prob}(4,114 < \sigma < 4,371) = 0,95$

Les intervalles de confiance cherchés sont donc :

$$\sigma^2 \in [16{,}91\ ;\ 19{,}09] \quad et \quad \sigma \in [4{,}112\ ;\ 4{,}371]$$

③ *Population approximativement normale*

On sait que

$$s^2 \to N(\sigma^2, s^2\sqrt{\frac{2}{n}}) \text{ or } T = \frac{18-\sigma^2}{2{,}427} \to N(0,1)$$

$$\text{Prob}\,(13{,}243 < \sigma^2 < 22{,}757) = 0{,}95$$

$$s \to N(\sigma, \frac{s}{\sqrt{2n}}) \Rightarrow \text{Prob}\,(3{,}682 < \sigma < 4{,}803) = 0{,}95$$

Les intervalles de confiance cherchés sont donc :

$$\sigma^2 \in [13{,}243\ ;\ 22{,}757] \quad et \quad \sigma \in [3{,}682\ ;\ 4{,}803]$$

7- On associe à p la variable aléatoire définie par

$$X = \begin{cases} 1 & \text{avec la probabilité } p \\ 0 & \text{avec la probabilité } q = 1 - p \end{cases}$$

On extrait un échantillon aléatoire simple d'effectif n et pour le $i^{ème}$ individu on note :

$$x_i = \begin{cases} 1 & \text{s'il possède le caractère} \\ 0 & \text{sinon} \end{cases}$$

La fréquence absolue f (nombre d'individus) possédant la caractéristique suit une *loi Binomiale B(n, p)* d'où :

$$\text{Prob}\,(\frac{f}{n} = \frac{k}{n}) = C_n^k p^k q^{n-k} \quad \text{avec } k = 0,\ 1,\ 2, \ldots,\ n$$

et $\quad E(\frac{f}{n}) = p \qquad V(\frac{f}{n}) = \frac{pq}{n}$

① *Méthode exacte*:

La connaissance de la distribution échantillonnée de la statistique $\frac{f}{n}$ permet de construire des intervalles de confiance aux coefficients 95 % et 99 % surtout pour les petites valeurs de n.

② *Approximation normale*

Le *théorème central limite* permet d'approximer la loi *binomiale B(n, p)* par une *loi normale* pour *n grand* (*np* et *nq supérieurs à 15*)

$$\frac{\frac{f}{n}-p}{\sqrt{\frac{pq}{n}}} \to N(0,1)$$

d'où Prob $(-t_{\alpha/2}\sqrt{\frac{pq}{n}} \le \frac{f}{n}-p \le t_{\alpha/2}\sqrt{\frac{pq}{n}}) = 1-\alpha$

En résolvant

$p = \frac{f}{n} \pm t_{\alpha/2}\sqrt{\frac{pq}{n}}$ par rapport à p, on obtient comme limite de l'intervalle de confiance :

$$\frac{n}{n+t^2}\left(\frac{f}{n}+\frac{t^2}{2n} \pm t\sqrt{\frac{f(1-f)}{n^3}+\frac{t^2}{4n^2}}\right) \quad \text{où} \quad t = t_{\alpha/2}$$

lorsque *n est très grand*, ces limites sont *approximativement égales* à :

$$\frac{f}{n} \pm t_{\alpha/2}\sqrt{\frac{f(1-f)}{n^3}}$$

8- ① X et Y suivent des *lois normales* de *variances connues*:

$$X \to N(m_1, \sigma_1) \qquad Y \to N(m_2, \sigma_2)$$

$$(\bar{x}-\bar{y}) - t^{\alpha/2}_{n_1+n_2-2}\sqrt{\frac{\sigma^2_1}{n_1}+\frac{\sigma^2_2}{n_2}} \le m_1-m_2 \le (\bar{x}-\bar{y}) + t^{\alpha/2}_{n_1+n_2-2}\sqrt{\frac{\sigma^2_1}{n_1}+\frac{\sigma^2_2}{n_2}}$$

② X et Y suivent des *lois normales* de *variances inconnues* mais *égales* :

$$(\bar{x}-\bar{y}) - t^{\alpha/2}_{n_1+n_2-2}\sqrt{\frac{n_1 s^2_1+n_2 s^2_2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le m_1-m_2$$

$$m_1-m_2 \le (\bar{x}-\bar{y}) + t^{\alpha/2}_{n_1+n_2-2}\sqrt{\frac{n_1 s^2_1+n_2 s^2_2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

③ X et Y de *lois quelconques* de *variances connues*

On utilise le *résultat trouvé* en ① pourvu que n_1 et n_2 *supérieures à 30*

④ X et Y de *lois quelconques* de *variances inconnues*

On procède comme en ③ en remplaçant σ_1 par s_1 et σ_2 par s_2.

Exercices

Exercice 1

① *Calcul de l'espérance mathématique et de la variance*

X	0,5	1	1,5	2	3	Σ
n_i	20	48	15	12	5	100
$n_i x_i$	10	48	22,5	24	15	119,5
$n_i x_i^2$	5	48	33,75	48	45	179,75

D'où $\bar{x} = \frac{119,5}{100} = 1,195$ et $s^2 = \frac{179,75}{100} - (1,195)^2 \# 0,3695$

On prend pour estimateur de l'espérance mathématique la moyenne observée soit $E(X) = 1,195$ et pour estimateur de la variance la quasi variance :

$$S^2 = \frac{ns^2}{n-1} = \frac{100 \times 0,3695}{99} = 0,37323$$

La taille de l'échantillon permet d'utiliser la loi normale centrée réduite d'où $t_{0,05} = 1,96$ ce qui donne pour intervalle :

$$[1,195 - 1,96\frac{0,6109}{10}; 1,195 + 1,96\frac{0,6109}{10}] = [1,0753; 1,3147]$$

② *Taille de l'échantillon*

On doit résoudre

$$t_\alpha \frac{S}{\sqrt{n}} = \frac{1,96 \times 0,6109}{\sqrt{n}} = 0,01 \text{ ; on obtient alors :}$$

$$n = (\frac{1,96 \times 0,6109}{0,01})^2 = 14336,8 \# 14337$$

Exercice 2

① *Calcul de l'espérance mathématique et de la variance*

X	0,85	0,88	0,90	0,92	0,95	1	Σ
n_i	1	2	4	3	3	3	16
$n_i x_i$	0,85	1,76	3,6	2,76	2,85	3	14,82
$n_i x_i^2$	0,7225	1,5488	3,24	2,5392	2,7075	3	13,758

D'où $\bar{x} = \frac{14,82}{16} \# 0,9263$ et $s^2 = \frac{13,758}{16} - (0,9263)^2 = 0,0018433$

On prend pour estimateur de l'espérance mathématique la moyenne observée soit $E(X) = 0,92625$ et pour estimateur de la variance la quasi variance :

$$S^2 = \frac{ns^2}{n-1} = \frac{16\times 0,0018433}{15} \# 0,001967$$

② La taille de l'échantillon étant petite, pour $\alpha = 0,02$; *degré de liberté 15* la lecture de la *table de Student* pour $t_\alpha(15) = 2,602$ ce qui donne :

$$[0,9263 - 2,602\frac{0,04435}{4}; 0,9263 + 2,602\frac{0,04435}{4}] = [0,89745;\ 0,9552]$$

Exercice 3

① *Intervalle de confiance pour σ_1^2 au niveau 90 %*

$$(n_1-1)\frac{S^2_1}{\sigma^2_1} \rightarrow \chi^2_{n_1-1}$$

Pour $\alpha = 0,1$; $n_1 = 9$ *et* $S^2_1 = 38$ on obtient :

$$\frac{\alpha}{2} = 0,05 \rightarrow \chi^2_{\alpha/2;\,\nu=8} = 2,73 \text{ et } 1-\frac{\alpha}{2} = 0,95 \rightarrow \chi^2_{1-\alpha/2;\,\nu=8} = 15,51$$

L'intervalle cherché est donc $\sigma_1^2 \in [19,60\ ;\ 111,335]$

② *Intervalle de confiance pour σ_2^2 au niveau 90 %*

Un raisonnement analogue au précédent nous donne

$$(n_2-1)\frac{S^2_2}{\sigma^2_2} \rightarrow \chi^2_{n_2-1}$$

Pour $\alpha = 0,1$; $n_2 = 6$ *et* $S^2_2 = 42$ on obtient :

$$\frac{\alpha}{2} = 0,05 \rightarrow \chi^2_{\alpha/2;\,\nu=5} = 1,15 \text{ et } 1-\frac{\alpha}{2} = 0,95 \rightarrow \chi^2_{1-\alpha/2;\,\nu=5} = 11,070$$

L'intervalle cherché est donc $\sigma_2^2 \in [18,970\ ;\ 182,609]$

③ Les deux intervalles ont une *grande partie commune*, ce qui suppose *l'égalité des deux variances*

④ *Intervalle de confiance pour $\frac{\sigma^2_1}{\sigma^2_2}$ au niveau 90 %*

$$(n_1-1)\frac{S^2_1}{\sigma^2_1} \rightarrow \chi^2_{n_1-1} \quad \text{et} \quad (n_2-1)\frac{S^2_2}{\sigma^2_2} \rightarrow \chi^2_{n_2-1}$$

donc $\dfrac{\dfrac{S^2_1}{\sigma^2_1}}{\dfrac{S^2_2}{\sigma^2_2}} \rightarrow F_{n_1-1;n_2-1}$ *F étant la loi de Fisher-Snedecor*

d'où $$\frac{1}{F_{n_1-1;n_2-1;\alpha/2}} \times \frac{S^2_1}{S^2_2} \leq \frac{\sigma^2_1}{\sigma^2_2} \leq \frac{1}{F_{n_1-1;n_2-1;1-\alpha/2}} \times \frac{S^2_1}{S^2_2}$$

En consultant la table de Fisher-Snedecor, on a :

$$F(8\ ;\ 5)_{0,95} = 4,82 \quad \text{et } F(5\ ;\ 8)_{0,95} = 3,69$$

$$\Rightarrow \frac{\sigma^2_1}{\sigma^2_2} \in [0,188\ ;\ 3,339]$$

⑤ *Intervalle de confiance pour $\mu_1 - \mu_2$ au niveau 90 %*

$$\mu_1 - \mu_2 \in [\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2;n_1+n_2-2} \times S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\ ]$$

où $S^2 = \dfrac{(n_1-1)S^2_1 + (n_2-1)S^2_2}{n_1+n_2-2}$ soit $S^2 = \dfrac{8\times38+5\times42}{13} = 39,538$

Comme $t_\alpha(13) = 1,771 \Rightarrow \mu_1 - \mu_2 \in [0,13\ ;\ 11,87]$

Exercice 4

Intervalle de confiance pour $\sigma_1^2 - \sigma_2^2$ et $\sigma_1 - \sigma_2$ au niveau 95 %

On sait que

$$S^2 \rightarrow N(\sigma^2, S^2\sqrt{\frac{2}{n}})$$

Pour déterminer l'intervalle de confiance, au niveau α, on utilise la distribution approchée de S^2 :

$$\text{Prob}\left(\left|\frac{S^2 - \sigma^2}{S^2\sqrt{\frac{2}{n}}}\right| \leq t_{\alpha/2}\right) = 1 - \alpha$$

et les bornes de l'intervalle sont pour σ^2

$$S^2 \pm t_{\alpha/2}S^2\sqrt{\frac{2}{n}} \qquad \text{Avec}$$

$T_1 = \dfrac{49-\sigma^2_1}{6,198} \rightarrow N(0,1)$ et $t_\alpha = 1,96$ $\qquad$ Prob$(36,852 < \sigma^2_1 < 61,148) = 0,95$

$T_2 = \dfrac{36-\sigma^2_2}{4,157} \rightarrow N(0,1)$ et $t_\alpha = 1,96$

$$\text{Prob}(27{,}852 < \sigma^2_2 < 44{,}148) = 0{,}95 \Rightarrow \sigma_1^2 - \sigma_2^2 \in [9\ ;\ 17]$$

On procède de la même manière pour le calcul de l'intervalle de la différence $\sigma_1 - \sigma_2$

$$S \to N(\sigma, S\sqrt{\frac{1}{2n}}) \quad \text{et} \quad S \pm t_{\alpha/2} S\sqrt{\frac{1}{2n}}$$

$$\text{Prob}(6{,}132 < \sigma_1 < 7{,}868) = 0{,}95 \quad \text{et} \quad \text{Prob}(5{,}887 < \sigma_2 < 6{,}113) = 0{,}95$$

$$\Rightarrow \sigma_1 - \sigma_2 \in [0{,}245\ ;\ 1{,}755]$$

Exercice 5

On rappelle :

- ✓ $\frac{ns^2}{\sigma^2} \to \chi^2_{n-1}$
- ✓ $\frac{S^2_1}{S^2_2} = \frac{\frac{n_1 s^2_1}{n_1 - 1}}{\frac{n_2 s^2_2}{n_2 - 1}} = F_{n_1-1;n_2-1}$ avec $\sigma^2_1 = \sigma^2_2$

loi de Fisher-Snedecor à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté

Il s'agit de déterminer 2 valeurs F' et F'' telles que

$$\text{Prob}(F_{n_1-1;n_2-1} < F') = 0{,}025 \quad \text{et} \quad \text{Prob}(F_{n_1-1;n_2-1} < F'') = 0{,}975$$

On lit dans la table $F' = 0{,}35$ et $F'' = 2{,}86$

D'où $\text{Prob}(0{,}35 < 3{,}125 \frac{\sigma_2^2}{\sigma_1^2} < 2{,}86) = 0{,}95 \quad \Rightarrow \quad \frac{\sigma_1^2}{\sigma_2^2} \in [1{,}09; 8{,}93]$

Exercice 6

① *Intervalle de confiance pour le temps d'attente moyen dans le service A à 90 %*

Comme l'écart-type est inconnu, on a :

$$\frac{\bar{x}_A - m_A}{\frac{s_A}{\sqrt{n_A}}} \in t_{n-1} \quad \text{d'où} \quad m_A \in [\bar{x}_A \pm t_A \frac{s_A}{\sqrt{n_A}}] \quad \text{avec} \quad t_A = 1{,}701 \quad \text{valeur lue}$$

sur la table de *Student* de *degré de liberté 28*. L'intervalle cherché est donc :

$$m_A \in [39{,}157\ ;\ 44{,}843]$$

② *Intervalle de confiance pour le temps d'attente moyen dans le service B à 90 %*

On procède de la même manière que précédemment, avec $t_B = 1,729$ valeur lue sur la table de *Student* de *degré de liberté 19*. L'intervalle cherché est donc : $m_B \in [32,293 ; 37,706]$

③ *Intervalle de confiance pour la différence*
En désignant par *1 l'hôpital A et par 2 l'hôpital B*

On sait que : $\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2;\alpha/2}\ S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

En utilisant la variable de *Student*, on obtient :

$$t_{n_1+n_2-2} \rightarrow \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n_1+n_2-2}}{n_1+n_2-2}}} = \frac{(\bar{x}_1-\bar{x}_2)-(m_1-m_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

avec $S^2 = \frac{(n_1-1)^2 s^2_1 + (n_2-1)^2 s^2_2}{n_1+n_2-2}$ qui est un estimateur sans biais de σ^2

on obtient, avec $t = 1,65$ $\Rightarrow m_1 - m_2 \in [3,043 ; 10,957]$

Exercice 7

① *Intervalle de confiance*

On sait que : $p = f \pm t_\alpha \sqrt{\frac{f(1-f)}{n}}$ et en utilisant le changement de variable T, on obtient avec $t_{0,05} = 1,96$ lue dans la table de la loi normale centrée réduite

$$p = 0,6 \pm 1,96\sqrt{\frac{0,6\times0,4}{100}} = 0,6 \pm 0,096 \Rightarrow p \in [0,504 ; 0,696]$$

② *Nombre d'étudiants à interroger*

$t_\alpha \sqrt{\frac{f(1-f)}{n}} = 0,01$ soit en remplaçant f la fréquence par sa valeur

$$n = 9220$$

③ *Il s'agit de déterminer* t_α

$$t_\alpha \sqrt{\frac{0,6\times0,4}{9220}} = 0,01 \Rightarrow t_\alpha = 1,960$$

Cette valeur correspond à un risque d'erreur de 0,95

Exercice 8

① *La taille de l'échantillon étant très grande n = 100, sa moyenne* suit la *loi normale* :

$$\text{Prob}\left(-t_{\alpha/2} \le \frac{\bar{x}-m}{\frac{\sigma}{\sqrt{n}}} \le t_{\alpha/2}\right)=1-\alpha$$

$t_{\alpha/2}$ est la valeur de la variable centrée réduite $T \to N(0,1)$ qui correspond au niveau α. Par lecture de la table de la loi normale centrée réduite, on trouve $t_{0,05} = 1,96$ d'où l'intervalle cherché $\Rightarrow m \in [3,81 ; 4,79]$

Le paramètre *m* a 95 chances sur 100 de prendre une valeur comprise entre *3,8 et 4,8*.

② *On connaît la variance empirique* s^2

On sait que :

$$S^2 = \frac{ns^2}{n-1} = \frac{100}{100-1}\,6,76 = 6,828$$

$\text{Prob}\left(-t_{\alpha/2} \le \frac{\bar{x}-m}{\frac{s}{\sqrt{n}}} \le t_{\alpha/2}\right)=1-\alpha$ et l'intervalle de confiance est donné par

$$\bar{x} \pm t_{n-1;\alpha/2}\,\frac{S}{\sqrt{n}}$$

Comme le degré de liberté est très grand la loi de Student tend vers la loi normale. Par lecture de cette dernière table, on trouve $t_{0,05} = 1,96$

$\Rightarrow m \in [3,79 ; 4,81]$ *celui de la question précédente.*

Exercice 9

Comme ρ est *inconnu* et que la *taille* de l'échantillon n'est *pas très grande*, on utilise la *transformation de Fisher Z'*. Or pour r = 0,3 correspond z' = 0,310 d'où l'intervalle

$$0,310 \pm 1,96\frac{1}{\sqrt{50-3}} = 0,31 \pm 0,286$$

Appliquons la *transformation inverse*

$z' = 0,024 \quad \to \quad r = 0,02$

$z' = 0,596 \quad \to \quad r = 0,53$

$\Rightarrow \rho \in [0,02 ; 0,53]$

Chapitre 5

TESTS PARAMETRIQUES

☛ *Objectifs*

⇨ Introduire la notion de test

⇨ étudier les tests d'hypothèses

⇨ présenter les tests de proportions, de moyennes, de variances

⇨ utiliser la loi de Student

☛ *Mots clés*

Test d'hypothèse ou test statistique – hypothèse nulle – hypothèse alternative – erreur de type I – erreur de type II – région de rejet région d'acceptation – test de signification – test bilatéral – test unilatéral – seuil de signification – coefficient de corrélation – transformation de Fisher.

I ***Principes généraux***

II ***Tests de proportions***

III ***Tests de moyennes***

IV ***Tests de variances***

V ***Test sur le coefficient de corrélation***

AUTO-EVALUATION & EXERCICES CORRIGES

Pour mieux comprendre la logique d'une décision d'un ***test statistique***, on cite souvent la logique d'une ***décision judiciaire*** en cour d'assises : une personne est ***innocente*** ou ***coupable.*** Si on accepte qu'elle est innocente, alors elle n'est pas coupable. Si on accepte qu'elle est coupable, elle n'est donc pas innocente.

Mais comment peut-on ***conclure à la culpabilité*** ? On arrive à cette conclusion, si et seulement si, il y a assez d'arguments pour douter de l'innocence.

Deux situations peuvent se présenter :

① Il y a *peu d'arguments*, on est dans le doute, on ne décide pas que la personne est coupable : la personne est déclarée *innocente*.

② Il y a *assez d'arguments* pour douter de son innocence: la personne est déclarée *coupable.*

Evidemment, on *peut se tromper* :

✧ Il y a *assez d'arguments* pour douter de l'innocence de la personne et la déclarer *coupable* bien qu'elle *soit innocente*.

✧ La masse d'arguments est jugée *insuffisante*, la personne est déclarée *innocente*, pour faute d'épreuves convaincantes, alors qu'en fait elle est *coupable.*

La logique d'un test *statistique est comparable* à celle d'une décision judiciaire. « On suppose *l'existence d'un phénomène* dans une population donnée, on *émet une hypothèse* et on *cherche à la vérifier*. »

Pour mieux comprendre les tests *d'hypothèses*, il est commode de faire encore appel à *deux exemples*.

Exemple 1

Une entreprise reçoit de son fournisseur un lot de lampes, fabriquées en série, respectant des normes déterminées. Considérons pour simplifier un seul caractère X (une seule norme) par exemple « la ***durée de vie des lampes***. » Chaque pièce du lot peut être classée en ***bonne*** ou ***mauvaise***, selon qu'elle satisfait ou non à la norme convenue. Le client peut exiger que toutes les lampes fournies doivent satisfaire à cette condition.

Dans la réalité, cette norme ne peut pas être définie exactement, *un intervalle de tolérance* est admis : le client accepte qu'une proportion p de lampes soit mauvaise.

Un accord peut être conclu entre le *fournisseur* et le *client* sur l'importance de cette proportion p : « le client ***accepte le lot*** si la proportion p est ***inférieure*** à un nombre p_0 et la ***refusera*** dans le cas ***contraire***. » Vu la nature du produit, la décision d'acceptation ou de refus sera prise sur un ***échantillon*** aléatoire.

A titre d'exemple, sur un lot de 100 pièces, la fabrication est :

- *acceptée* si *moins de 6 pièces* sont *défectueuses*
- *refusée* dans le *cas contraire*.

La proportion retenue (p_0 = 6%) n'est pas la proportion exacte dans l'ensemble de la fabrication. Elle s'en écarte, plus au moins, selon le *hasard de l'échantillonnage*. *L'acceptation (ou le refus)* des lots est donc un *caractère aléatoire* dont il faut *étudier les fluctuations.*

Exemple 2

Une entreprise lance un nouveau médicament sur le marché. Elle prétend que son produit est *efficace* dans 75% des cas. Un laboratoire de contrôle teste le produit, à partir d'un échantillon aléatoire de taille n =100. Il décide d'***accepter*** le pourcentage avancé, si le nombre de cas favorable est supérieur ou égal à 70 et de ***le rejeter*** dans le cas où le nombre serait *inférieur ou égal* à 69.

Deux décisions sont donc possibles :

① *Rejeter* le pourcentage avancé par l'entreprise alors qu'il est *exact* : c'est le *risque du vendeur*

② *Accepter* le pourcentage avancé alors qu'il est *faux* : c'est le *risque de l'acheteur.*

Il s'agira de *mesurer les deux risques.*

Comme dans les deux exemples précédents, on considère une population ***P*** dont les éléments possèdent un *caractère mesurable X* dont la valeur de *l'un des paramètres* θ (proportion, moyenne, écart- type...) est ***inconnue.*** Dans les chapitres déjà étudiés, on a estimé ces paramètres à partir des *théories de l'échantillonnage et de l'estimation.*

Dans la *théorie des tests,* le principe demeurant le même, on est amené, à partir *d'informations incomplètes* sur la population ***P,*** *d'admettre une hypothèse* puis de chercher à la *vérifier.*

Tout en étudiant les caractéristiques de l'échantillon aléatoire, tiré de cette population ***P***, on cherche à *porter un jugement sur la validité* de l'hypothèse *avancée.* Une telle hypothèse est dite :

- *paramétrique* lorsqu'elle concerne la valeur des *paramètres*
- *non- paramétrique* lorsqu'elle porte sur la *forme de la distribution.*

Nous développerons précisément le cas paramétrique, en utilisant l'approche de *Bayes* et surtout celle de *Neyman.* Comme pour l'étude de l'estimation et de l'échantillonnage, les développements mathématiques approfondis seront peu utilisés.

I - PRINCIPES GENERAUX

I - 1 TERMINOLOGIE

▪ On appelle ***test d'hypothèse*** ou ***test statistique,*** *une démarche* dont le but est de fournir une *règle de décision* permettant, sur la base des résultats *d'échantillon,* de faire un *choix entre deux hypothèses* statistiques.

▪ Ayant observé un *échantillon de taille* ***n*** $(X_1,\ldots, X_n)$, soit H_0 *l'hypothèse spécifique retenue* relative à la variable X. *Deux décisions* sont possibles :

① Considérer H_0 comme *vraie* $\Rightarrow$ *on l'accepte*

② Considérer H_0 comme *fausse* $\Rightarrow$ on *la rejette*

▪ A *ces deux décisions* correspondent *deux régions complémentaires* de R^n :

① Une région de *rejet de* H_0 (W) appelée *région critique* où *l'hypothèse* n'est *pas retenue.*

② Une région de *non-rejet* ou *d'acceptation* de H_0 où *l'hypothèse* est *retenue.*

▪ C'est la *règle de décision générale* ou *test statistique*. L'un des aspects importants est *la détermination de la région de rejet.*

▪ Si H_0 *est vraie,* on peut calculer pour toute région ***W,*** la probabilité de *rejeter* H_0 *à tort* et la désigner par α

➢ α = ***Prob (rejeter*** H_0/ H_0 ***vraie)*** appelée ***erreur de type I*** (*condamner* une personne alors qu'elle est *innocente*)

➢ D'autre part, on peut commettre une *erreur d'un autre type* (***erreur de type II)*** ne *pas rejeter* H_0 alors qu'elle *est fausse* (déclarer quelqu'un *innocent* alors qu'il est *coupable*.) La probabilité de *ne pas rejeter* H_0 ***alors*** *qu'elle est fausse est appelée* :

$\beta = 1 - \alpha$ = Prob (ne pas ***rejeter*** H_0/ H_0 ***fausse)***

Pour la *déterminer*, il faut définir une autre *hypothèse alternative* H_1 correspondant à la probabilité $\beta = 1 - \alpha$

■ Au seuil de α, on peut calculer β et son complément $(1 - \beta)$ appelé *puissance du test.*

■ Un *test statistique* sera toujours *un test entre deux hypothèses* H_0 ***et*** H_1 ***jouant des rôles symétriques.***

Remarques

☞ H_0 ***et*** H_1 ne peuvent pas être *vraies simultanément.*

On *soumet* H_0 *au test* qu'on considère comme *vrai*. Si le test conduit, d'après l'étude de l'échantillon, au *rejet de* H_0 alors H_1 est considérée comme *vraisemblable*.

☞ On peut schématiser la règle de décision comme suit :

Conclusion du test

Réalité / ***D***écision	H_0	H_1
H_0	$1 - \alpha$	β
H_1	α	$1 - \beta$

I - 2 METHODES

Construire un test, c'est définir :

* *l'espace des observations* $\mathbf{R}^n$,
* l'ensemble Θ des *valeurs possibles du paramètre* θ,
* *les deux décisions possibles.*

A un *problème donné*, correspondent plusieurs *règles de décisions* : Deux méthodes sont utilisées : la *méthode de Bayes* et *la méthode de Neyman.*

I- 2. 1 Méthode de Bayes

① **Analyse**

On suppose qu'il existe deux valeurs possibles θ_0 *et* θ_1 dans l'espace des solutions Θ. On leur attribue des *probabilités à priori.* On calcule ensuite les *coûts associés à chacune des décisions pour chaque valeur du paramètre.* L'observation de l'échantillon permet, en utilisant le théorème de Bayes, de calculer les probabilités π_0 et π_1 *à posteriori* des décisions D_0 et D_1 relatives aux hypothèses H_0 ***et*** H_1.

② ***Tableau des décisions***

Décision / *Paramètre*	D_0	D_1	Probabilité à priori	Probabilité à posteriori
θ_0	C_{00}	C_{01}	p_0	π_0
θ_1	C_{10}	C_{11}	$p_1 = 1 - p_0$	π_1

Les décisions D_0 et D_1 sont des *variables aléatoires.*

③ ***Décision***

La décision *optimale* est celle qui *minimise l'espérance mathématique du coût de chaque décision.*

Remarque

☞ Au lieu des coûts, on utilise les avantages liés aux décisions D_0 et D_1. La décision *optimale* est celle qui *maximise l'espérance mathématique du gain pour chaque décision.*

I- 2. 2 Méthode de Neyman

La méthode de choix de la région *(W)* que nous avons exposée précédemment, est celle de *Neyman et Pearson.* Elle est la plus utilisée. Elle repose sur le *principe suivant* : « α *étant fixé, choisir la région* (W_0) *de façon à maximiser la puissance du test. » Plusieurs cas à examiner* :

a) **Deux hypothèses simples**

① ***Analyse***

On se propose de choisir *deux hypothèses* sur les valeurs numériques du paramètre inconnu θ de la distribution X de *densité de probabilité* $f(x, \theta)$.

Connaissant la *vraisemblance L(x, θ)* de l'échantillon, il s'agit de tester :

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$$

② ***Décision***

Comme α = Prob (W / H_0 ***vraie***) = $\int_w L(x, \theta_0)\, dx$

et la puissance du test : *1* - β = Prob (W / H_1 ***vraie***) = $\int_w L(x, \theta_1) dx$

Cette *puissance* peut être interprétée comme l'espérance de la variable aléatoire :

$$\frac{L(x, \theta)}{L(x, \theta_0)}$$

En effet 1- β = $\int_w \frac{L(x, \theta)}{L(x, \theta_0)} L(x, \theta_0)\, dx$ et maximiser 1- β revient à choisir la région W, de *probabilité* α, *sous hypothèse* H_0 telle que le rapport des vraisemblances $\frac{L(x, \theta)}{L(x, \theta_0)}$ soit *maximum*. On obtient ainsi un *procédé général de détermination* de la *région critique* :

$$\frac{L(x, \theta)}{L(x, \theta_0)} > k_\alpha$$

b) **Une hypothèse simple et une hypothèse multiple**

Donnons *un exemple* :

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta < \theta_0 \end{cases} \qquad \begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

L'hypothèse H_0 est dite *simple* et l'hypothèse H_1 *multiple* ou *composite* parce qu'elle est composée d'un ensemble de valeurs de θ. Pour chaque valeur particulière, on peut calculer 1- β.

c) **Test U.M.P.**

Un test est dit *U..M..P.* (en anglais uniformly most powerfull) ou *U.P.P* (uniformément le plus puissant) si, quelle que soit la valeur de θ appartenant à l'hypothèse alternative H_1, sa *puissance est supérieure* à la puissance de *tout autre test*.

Remarques

☞ Pour ce test, la région critique déterminée est telle que β est minimum et la puissance du test maximum quelle que soit l'hypothèse alternative.

☞ Ce test n'existe pas toujours, c'est le cas de l'hypothèse :

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Dans ce cas *la région critique* est *déterminée comme suit* :

* On *choisit* une région d'acceptation unique de H_0 formée de la partie commune de toutes les régions d'acceptation relatives aux différentes valeurs possibles de H_1

* Cette région est fonction de H_0 et de α. Elle est indépendante de H_1. Son aire est égale à $1 - \alpha$.

I - 3 APPLICATIONS

✧ L'hypothèse de base étant celle dont on considère que le rejet, à tort, aura de lourdes conséquences.

① **L'hypothèse nulle** H_0

On fixe à priori un paramètre de la population à une *valeur* particulière. On admet que seul *le hasard a provoqué la variation* des résultats obtenus.

② **L'hypothèse alternative** H_1 est la *contre hypothèse* c'est à dire n'importe qu'elle autre hypothèse. On admet que la variation n'est pas due uniquement au hasard, mais à *une différence des populations* concernées.

✧ Pratiquement on utilise une démarche qui permettrait, à long terme, de *rejeter à tort une hypothèse nulle vraie* dans une *faible proportion* de cas. La conclusion déduite des résultats de l'échantillon suivant la règle adoptée aura un caractère probabiliste. On décide tout en sachant à l'avance qu'il y a un certain *risque α que la décision ne soit pas la bonne*. Ce risque est donné par *le seuil de signification* du test c'est à dire la réponse donnée à la question suivante :

« Est -ce que le hasard ou les variations d'échantillonnage peuvent donner une explication vraisemblable de la différence entre le résultat observé et la valeur retenue dans H_0 ? »

Deux réponses possibles :

① Une *réponse affirmative* : le *résultat* observé est *compatible* avec l'hypothèse retenue : H_0 est *vraie* $\Rightarrow$ on ne la *rejette pas* $\Leftrightarrow$ le test est *statistiquement non significatif*

② Une *réponse négative* : le *résultat* observé *n'est pas compatible* : H_0 est *fausse* $\Rightarrow$ on la *rejette* $\Leftrightarrow$ le test est *statistiquement significatif.*

Le tableau suivant résume les termes souvent utilisés. Les *lignes* ayant la *même signification* :

Statistiquement non significatif	ne *pas rejeter* H_0	*valeur* observée *compatible*	Les variations expliquent la différence constatée, mais ne prouvent pas H_0
Statistiquement significatif	*rejeter* H_0	*valeur* observée non *compatible*	Les variations n'expliquent pas la différence constatée

Remarques

☞ Un test de signification se rapporte toujours à l'hypothèse nulle, celle qui fait l'objet du test.

☞ Le résultat du test ne se réduit pas aux affirmations suivantes :
« *non l'hypothèse n'est pas vérifiée* » ou « *oui l'hypothèse est vraie* .»

Il peut nous indiquer le choix entre :
« rejeter l'hypothèse avec un risque donné » ou « ne pas la rejeter avec le même risque. »

☞ Les conclusions du test ne sont pas symétriques.

☞ La puissance du test mesure sa capacité à mettre en évidence la différence entre la valeur observée et la valeur prédite.

☞ On utilise généralement les seuils suivants :

- $\alpha = 0{,}05 \Leftrightarrow$ prendre un risque de 5 % de rejeter l'hypothèse H_0 alors qu'elle est vraie $\Leftrightarrow$ les tests effectués à ce seuil conduiront une fois sur vingt au rejet de H_0 alors qu'elle est vraie.
- $\alpha = 0{,}01$; $\alpha = 0{,}001$

☞ Un résultat *significatif à un niveau* l'est également à tous les niveaux de signification plus grands : significatif à 0,1 % ⇒ significatif aux niveaux supérieurs : 1 % ; 5 % ; 10 %...

Ayant *défini les tests d'hypothèses* et *souligné leurs propriétés* les plus importantes, on se propose de donner la marche à suivre pour leur *construction*. On s'intéressera particulièrement aux tests usuels : tests sur les proportions, les moyennes, les variances, le coefficient de corrélation.

II - TESTS DE PROPORTIONS

II - 1 TEST SUR UNE PROPORTION

Selon les hypothèses retenues, on *distingue deux régions* :

① une *région critique bilatérale* ou *test bilatéral* (composé de deux intervalles) ;

② une *région critique unilatérale* ou *test unilatéral* (composé d'un seul intervalle).

II - 1. 1 Test bilatéral

✧ On considère une population $\boldsymbol{P}$ et un échantillon aléatoire prélevé avec remplacement dans cette population. On suppose *n assez grand* pour pouvoir utiliser *l'approximation* de la *loi binomiale* par une *loi normale*.

On teste *l'hypothèse nulle* : $H_0 : \boldsymbol{p = p_0}$

par rapport à *l'hypothèse alternative* $H_1 : \boldsymbol{p \neq p_0}$

Faut-il rappeler qu'on ne teste que les hypothèses qui sont sujettes à des doutes et qu'on prend pour fausses. Par conséquent l'acceptation est sujette à des réserves, car d'autres tests plus puissants avec prélèvements d'autres échantillons sont nécessaires. La technique utilisée peut se schématiser comme suit :

- *Si on considère que l'hypothèse est vraie, son acceptation ne constituera pas une raison en sa faveur*
- *Par contre son rejet est considéré comme un démenti à la proposition avancée*.

Il s'agit de prendre *une décision* au sujet de la valeur de la proportion p inconnue, d'éléments de la population *possédant le caractère A*. Les distributions d'échantillonnage d'une proportion traitées dans les chapitres

précédents, indique que la meilleure estimation de p est la fréquence relative observée $\frac{k}{n}$. Comparons cette fréquence à une proportion hypothétique $\boldsymbol{p_0}$

En supposant H_0 ***vraie***, $\frac{k}{n}$ possède une distribution $B(n, p)$, qu'on peut approximer par une *loi normale* de moyenne $\boldsymbol{p_0}$ et d'écart–type $\sqrt{\frac{p_0 q_0}{n}}$, on a alors : $\frac{k}{n} \rightarrow \mathrm{N}(p_0, \sqrt{\frac{p_0 q_0}{n}})$

L'hypothèse H_0 est soit

- *acceptée* si la valeur observée est très peu différente de $\boldsymbol{p_0}$
- *rejetée* sinon

On commettra alors soit

- une *erreur de première espèce* en rejetant $\boldsymbol{p = p_0}$ alors que H_0 ***est vraie***
- une *erreur de seconde espèce* en retenant $\boldsymbol{p = p_0}$ alors que H_0 ***est fausse***

① ***Analyse***

Soit un niveau de probabilité fixé α de commettre l'*erreur de première espèce*. On a : $\frac{k}{n} \rightarrow \mathrm{N}(p_0, \sqrt{\frac{p_0 q_0}{n}})$ et la variable centrée réduite $T = \dfrac{\frac{k}{n} - \mathrm{p_0}}{\sqrt{\frac{\mathrm{p_0 q_0}}{n}}} \rightarrow N(0, 1)$

Or Prob $(-t_{\alpha/2} \leq \dfrac{\frac{k}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \leq + t_{\alpha/2}) = 1 - \alpha$ où $t_{\alpha/2}$ est une *valeur lue* dans la table de la *loi normale centrée réduite N(0, 1)*.

Si H_0 ***est vraie***, il y a une probabilité $1 - \alpha$ pour que T appartienne à l'intervalle$[-t_{\alpha/2}; t_{\alpha/2}]$

$$\text{Prob}\left(p - t_{\alpha/2}\sqrt{\frac{pq}{n}} \le p \le p + t_{\alpha/2}\sqrt{\frac{pq}{n}}\right) = 1 - \alpha$$

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée* si $\left| T = \dfrac{\frac{k}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \right| \le t_{\alpha/2}$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon***

La probabilité de commettre l'*erreur de première espèce* est la probabilité de rejeter l'hypothèse alors qu'elle est vraie :

$$\text{Prob}[(T < -t_{\alpha/2}) \cup (T > t_{\alpha/2})]$$

Cette probabilité est égale à $1 - \text{Prob}(|T| < t_{\alpha/2}) = 1 - (1 - \alpha) = \alpha$

Remarques

☞ Rappelons qu'accepter une hypothèse ne prouve pas qu'elle soit vraie, mais qu'il n'y a pas de désaccord entre la théorie et les observations, pour le test utilisé et au seuil retenu.

☞ Rejeter H_0 signifie que $\frac{k}{n}$ est différent de p_0 ou que la *différence* $\frac{k}{n} - p_0$ est *significative* et n'est pas *imputable au hasard.*

Exemple

On teste l'efficacité d'un nouveau médicament, au niveau de 0,05, H_0 ***vraie*** : $p = p_0 = 0{,}75$ par rapport à H_1 : $p \neq p_0$ au moyen d'un échantillon de taille $n = 100$ et $k = 70$

◆ $\frac{k}{n} = 0{,}7$ $\quad \alpha = 0{,}05$ $\quad t_{\alpha/2} = 1{,}96$ $\quad \dfrac{\frac{k}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = -1{,}15$

La valeur absolue n'est pas inférieure à 1,96 $\Rightarrow$ on ne rejette pas H_0.

II - 1. 2 Test unilatéral

Il s'agit de tester par exemple *l'hypothèse nulle* H_0 ***vraie*** : $\boldsymbol{p \geq p_0}$ par rapport à l'hypothèse alternative H_1 : $\mathbf{p < p_0}$

① ***Analyse***

Un raisonnement analogue au précédent conduit à considérer que $\frac{k}{n}$ possède une distribution approximativement normale :

$$\frac{k}{n} \rightarrow \mathrm{N}\left(p_0, \sqrt{\frac{p_0 q_0}{n}}\right)$$

Intuitivement, on *rejette l'hypothèse* si $\frac{k}{n}$ est *trop inférieur* à $\boldsymbol{p_0}$

Au niveau de probabilité α, arbitrairement fixé, correspond la valeur t_α de la variable centrée réduite T, telle que la fonction de répartition :

$G(t_\alpha) = 1 - \alpha$

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée* si $T = \dfrac{\frac{k}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} > -t_\alpha$

on ne rejette pas l'hypothèse : $\boldsymbol{p \geq p_0}$

✧ ***L'hypothèse H_0*** est *rejetée* si $T = \dfrac{\frac{k}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \leq -t_\alpha$ pour $\boldsymbol{p = p_0}$ et à plus forte raison pour $\boldsymbol{p \geq p_0}$

Exemple

On se propose de prouver éventuellement, que le nouveau médicament n'est pas aussi efficace que le laboratoire ne le prétend et que l'efficacité n'est pas supérieure à 75 %.

Tester *l'hypothèse nulle*

$$\begin{cases} H_0 : \boldsymbol{p \geq 0,75} \text{ par rapport à } \textit{l'hypothèse alternative} \\ H_1 : \boldsymbol{p < 0,75} \end{cases}$$

◆ $\frac{k}{n} = 0,7$ $\quad \alpha = 0,05$ $\quad t_{\alpha/2} = 1,645$ $\quad \dfrac{\frac{k}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = -1,155$

Cette valeur n'est pas inférieure à $-\ 1,645 \Rightarrow$ on ne rejette pas $\boldsymbol{H_0}$, l'échantillon ne permet pas de démentir l'affirmation du laboratoire.

◆ Les limites de l'intervalle d'acceptation du test bilatéral étant [-1,96 ; 1,96], la valeur de la variable – 1,155 étant plus proche de la zone de rejet (-1,654), on conclut que le test bilatéral est plus sévère, il conduit au rejet.

II - 2 TEST DE COMPARAISONS

❑ On considère deux populations $\boldsymbol{P_1}$ et $\boldsymbol{P_2}$ partagées en deux classes sur la base du critère « avoir le caractère A »

❑ $\boldsymbol{P_1}$ admet une distribution $B(1, p_1)$ et $\boldsymbol{P_2}$ admet une distribution $B(1, p_2)$; p_1 *et* p_2 étant les deux *proportions inconnues qu'on cherche à comparer*.

❑ On prélève *avec remplacement* et de *manière indépendante deux échantillons aléatoires* d'effectifs n_1 et n_2.

❑ Les *deux échantillons* sont *supposés grands* pour pouvoir utiliser l'approximation de la *loi binomiale* par la *loi normale*.

❑ On se propose de comparer p_1 *et* p_2 à partir des deux échantillons.

On utilisera $\frac{k_1}{n_1}$ comme estimation de p_1 et $\frac{k_2}{n_2}$ comme estimation de p_2

Les variables $\frac{k_1}{n_1}$ et $\frac{k_2}{n_2}$ admettent des distributions binomiales $B(p_1, n_1)$ et $B(p_2, n_2)$.

❑ On *distingue deux régions* :

II - 2. 1 Test bilatéral

On teste *l'hypothèse nulle* : $\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases}$

par rapport à *l'hypothèse alternative*

① ***Analyse***

Les deux variables étant indépendantes et en utilisant la stabilité de la loi normale, on peut écrire que la distribution de $\frac{k_1}{n_1} - \frac{k_2}{n_2}$ est *approximativement normale* :

$$\frac{k_1}{n_1} - \frac{k_2}{n_2} \rightarrow N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

Soit p la valeur commune inconnue de p_1 *et* p_2 et $q = 1 - p$

$$\frac{k_1}{n_1} - \frac{k_2}{n_2} \rightarrow N\left(0, \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Les deux populations étant identiques, nous pouvons considérer que :

* les deux échantillons sont prélevés dans *une même population*
* qu'ils constituent *un seul échantillon*
* que la *meilleure estimation* de $p = p_1 + p_2$ est donnée par :

$$\hat{p} = \frac{k_1 + k_2}{n_1 + n_2} \text{ et } \hat{q} = 1 - \hat{p} \quad ; \quad \frac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \rightarrow N(0,1)$$

Au niveau de *probabilité* α, correspond la valeur $t_{\alpha/2}$ de la variable normale centrée réduite telle que $\text{Prob}(|T| < t_{\alpha/2}) = 1 - \alpha$ et approximativement :

$$\text{Prob}\left(\frac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| \leq t_{\alpha/2}) = 1 - \alpha$$

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée* si $\left| \dfrac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_1}+\frac{1}{n_2})}} \right| \leq t_{\alpha/2}$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon.***

Intuitivement, on *rejette l'hypothèse* si $\frac{k_1}{n_1}$ et $\frac{k_2}{n_2}$ ne sont pas voisins ou quand leur *différence est trop grande*.

II - 2. 2 Test unilatéral

On teste, par exemple, *l'hypothèse nulle* H_0 : $p_1 = p$ par rapport à *l'hypothèse alternative* H_1 : $p_1 < p_2$

① ***Analyse***

$$\text{Prob}(|T| > t_{\alpha/2}) = \alpha$$

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée* si $\left| \dfrac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_1}+\frac{1}{n_2})}} \right| \leq t_{\alpha/2}$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon.***

Remarques

☞ ***Echantillons appariés* :TEST DE MCNEMAR**

① ***Analyse***

On administre *deux traitements A* et *B* aux mêmes individus. On compare les pourcentages de succès, les résultats étant notés (+) et (-). On construit le *tableau de contingence* suivant :

Echantillon 1 / *Echantillon 2*	+	-
+	*y*	***s***
-	***r***	*x*

En appelant :
y « *succès dans les 2 échantillons* »(+)
r « *succès dans l'échantillon 1* » (+) et « *échec dans l'échantillon 2* »(-)
s « *succès dans l'échantillon 2* » (+) et « *échec dans l'échantillon 1* »(-)
x « *échec dans les 2 échantillons* »(-)

Il s'agit de *comparer le nombre de paires de réponses* :

$\frac{r}{r+s}$ à $\frac{1}{2}$ ou r à $\frac{r+s}{2}$; soit : $T = \frac{|r-s|}{\sqrt{r+s}}$

② ***Décision***

Les résultats *y et x ne donnent aucune information sur la différence* des deux échantillons, ils ne *seront pas utilisés*.
La variable T suit une *distribution approximativement normale*

✧ ***L'hypothèse H_0*** ($r = s$) est *acceptée* si $t_{observée} \leq t_\alpha$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon.***

Ce test est appelé test de MCNEMAR. Il est valable lorsque le nombre de paires considérées est supérieur ou égal à 10

III - TESTS DE MOYENNES

On considère la population ***P*** retenue pour l'estimation d'une moyenne. *On suppose* que cette moyenne est égale à une valeur particulière m_0. On se propose de résumer les différents tests qui peuvent se présenter et donner leurs régions critiques.

III - 1 TEST SUR UNE MOYENNE

On teste *l'hypothèse nulle* : $H_0 : m = m_0$
par rapport à *l'hypothèse alternative* $H_1 : m \neq m_0$

Selon la distribution de la population, on vérifie les hypothèses, selon que la variance est donnée ou estimée par l'échantillon.

III - 1. 1 Distribution quelconque n ≥ 30

a) **σ connu**

① ***Analyse***

✧ $\bar{x}$ est l'estimateur de m

✧ $\bar{x}$ suit la distribution $N(m_0, \frac{\sigma}{\sqrt{n}})$

✧ On a approximativement: $\text{Prob}(-t_{\alpha/2} \leq \frac{\bar{x}-m_0}{\frac{\sigma}{\sqrt{n}}} \leq t_{\alpha/2}) = 1-\alpha$

$t_{\alpha/2}$ est la valeur de la variable centrée réduite $T \to N(0,1)$ qui correspond au niveau α.

② ***Décision***

✧ ***L' hypothèse H_0*** est *acceptée* si $\left| \frac{\bar{x}-m_0}{\frac{\sigma}{\sqrt{n}}} \right| \leq t_{\alpha/2}$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon.***

Intuitivement, on *rejette l'hypothèse* quand $\bar{x}$ est trop éloignée de m_0

b) **σ inconnu**

Pour construire un test sur m, lorsque la variance est inconnue, on estime l'écart-type à partir de l'échantillon. Soit S l'écart-type non biaisé de σ. Quelle que soit la distribution de la population, si la taille de l'échantillon est assez grande, la statistique : $\frac{\bar{x}-m}{\frac{S}{\sqrt{n-1}}}$ suit approximativement une loi normale centrée réduite.

① ***Analyse***

✧ $\bar{x}$ est l'estimateur de m

✧ $\bar{x}$ suit la distribution $N(m_0, \frac{S}{\sqrt{n-1}})$

✧ On a approximativement: $\text{Prob}(-t_{\alpha/2} \leq \frac{\bar{x}-m_0}{\frac{S}{\sqrt{n-1}}} \leq t_{\alpha/2}) = 1-\alpha$

$t_{\alpha/2}$ est la valeur de la variable centrée réduite $T \to N(0,1)$ qui correspond *au niveau* α.

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée* si $\left| \dfrac{\bar{x}-m_0}{\dfrac{S}{\sqrt{n-1}}} \right| \leq t_{\alpha/2}$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon.***

III- 1. 2 Distribution normale

On s'intéresse au cas où σ **est inconnu**

① ***Analyse***

✧ On sait que la statistique :

$\dfrac{\bar{x}-m}{\dfrac{S}{\sqrt{n}}}$ suit une *loi de Student* de *(n – 1) degrés de liberté*

✧ $\text{Prob}\left(-t_{n-1;\alpha/2} \leq \dfrac{\bar{x}-m_0}{\dfrac{S}{\sqrt{n}}} \leq t_{n-1;\alpha/2}\right) = 1-\alpha$

② ***Décision***

✧ ***L' hypothèse H_0*** est *acceptée* si $\left| \dfrac{\bar{x}-m_0}{\dfrac{S}{\sqrt{n}}} \right| \leq t_{n-1;\alpha/2}$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon.***

III - 2 TEST DE COMPARAISONS

❑ Pour une seule population ***P***, nous appelons ***m*** la moyenne, σ^2 la variance, μ_r le moment centré d'ordre *r*. Nous prélevons dans cette population un échantillon aléatoire de taille ***n***.

❑ Pour deux populations, on affecte respectivement les notations précédentes par les indices 1 et 2 :

$P_1 \Rightarrow m_1$; σ^2_1 ; ***moments centrés d'ordre quatre :*** $\mu_4^{(1)}$

Echantillon n_1 ; s^2_1 ; S^2_1

$P_2 \Rightarrow m_2$; σ^2_2 ; ***moments centrés d'ordre quatre :*** $\mu_4^{(2)}$

Echantillon n_2 ; s^2_2 ; S^2_2

Il s'agit dans ce qui suit de comparer les deux moyennes.

On ***teste alors les hypothèses*** :

par rapport à $\begin{cases} H_0 : m_1 = m_2 \\ H_1 : m_1 \neq m_2 \end{cases}$

dans les cas d'une population *quelconque* et d'une population *normale*.

III - 2. 1 *Population quelconque*

a) ***σ_1 et σ_2 connus***

① ***Analyse***

✧ La variable différence $\bar{x}_1 - \bar{x}_2 \rightarrow N\left(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$

✧ Puisque $m_1 = m_2$ on définit la variable *normale centrée réduite* :

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}} \rightarrow N(0,1)$$

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée* si $\left| \bar{x}_1 - \bar{x}_2 \right| \leq t_{\alpha/2} \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon***

b) ***σ_1 et σ_2 inconnus***

① ***Analyse***

✧ On remplace les écarts types des formules précédentes par les variances non biaisées des échantillons respectifs S^2_1 et S^2_2

② ***Décision***

✧ ***L' hypothèse H_0*** est *acceptée* si $\left| \bar{x}_1 - \bar{x}_2 \right| \leq t_{\alpha/2} \sqrt{\frac{S^2_1}{n_1} + \frac{S^2_2}{n_2}}$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon***

III - 2. 2 *Population normale*

On s'intéresse au cas où ***σ_1 et σ_2 inconnus***

✧ Supposons H_0 vraie $\sigma_1 = \sigma_2$

① ***Analyse***

L'intervalle de confiance pour $(m_1 - m_2)$, au *niveau* α, *a pour bornes* : $\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2;\alpha/2} \; S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

② ***Décision***

✧ ***L' hypothèse H_0*** est *acceptée* si

$$\left| \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \right| \leq t_{n_1+n_2-2;\alpha/2}$$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon***.

IV - TESTS DE VARIANCES

IV - 1 TEST SUR UNE VARIANCE

✓Soit ***P*** une population de moyenne ***m*** et d'écart-type σ.

✓On utilise de préférence les *estimateurs sans biais* de la variance

$$S^2 = \frac{ns^2}{n-1}$$

✓ La taille de l'échantillon n est *grande* $n \geq 50$

✓ Il s'agit dans ce qui suit de ***tester les hypothèses $H_0 : \sigma^2 = \sigma^2_0$*** par rapport à $H_1 : \sigma^2 \neq \sigma^2_0$ dans les cas où la population est *quelconque, approximativement normale* et *normale*.

IV - 1. 1 *Population quelconque*

✧ Supposons H_0 vraie $\sigma^2 = \sigma^2_0$

① ***Analyse***

La variance et l'écart-type suivent respectivement :

$$S^2 \to N(\sigma_0{}^2, \sqrt{\frac{\mu_4 - \sigma_0{}^4}{n}}) \qquad \text{et} \qquad S \to N(\sigma_0, \sqrt{\frac{\mu_4 - \sigma_0{}^4}{4n\sigma_0}})$$

μ_4 remplacé par son estimation m_4 (moment d'ordre quatre), on obtient les variables centrées réduites:

$$\frac{S^2 - \sigma_0^2}{\sqrt{\frac{m_4 - \sigma_0^4}{n}}} \quad \text{et} \quad \frac{S - \sigma_0}{\sqrt{\frac{m_4 - \sigma_0^4}{4n\sigma^2_0}}}$$

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée* si

$$\left| \frac{S^2 - \sigma_0^2}{\sqrt{\frac{m_4 - \sigma_0^4}{n}}} \right| \leq t_{\alpha/2}$$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon***.

✧ *L'intervalle d'acceptation* correspondant à H_0 : $\sigma = \sigma_0$ est

$$\frac{S - \sigma_0}{\sqrt{\frac{m_4 - \sigma_0^4}{4n\,\sigma^2_0}}} \leq t_{\alpha/2}$$

IV - 1. 2 ***Population approximativement normale***

✧ Supposons H_0 vraie $\sigma^2 = \sigma^2_0$

① ***Analyse***

On estime la variance par celle de l'échantillon

$S^2 \rightarrow N(\sigma_0^2, \sigma_0^2\sqrt{\frac{2}{n}})$ de même pour l'écart-tpe soit $S \rightarrow N(\sigma_0, \sigma_0\sqrt{\frac{1}{2n}})$

Il en résulte pour la variance

$$\text{Prob}\left(\left| \frac{S^2 - \sigma^2_0}{\sigma^2_0\sqrt{\frac{2}{n}}} \right| \leq t_{\alpha/2} \right) = 1 - \alpha$$

② ***Décision***

✧ ***L' hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si la valeur de S^2 appartient à l'intervalle :

$$\sigma_0^2 \pm t_{\alpha/2}\,\sigma_0^2\sqrt{\frac{2}{n}}$$

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***.

✧ *L'intervalle d'acceptation* correspondant à H_0 : $\sigma = \sigma_0$ est

$$\sigma_0 \pm t_{\alpha/2}\,\sigma_0\sqrt{\frac{2}{n}}$$

IV - 1. 3 ***Population normale***

✧ Supposons H_0 vraie $\sigma^2 = \sigma^2_0$

Les deux *paramètres sont inconnus*.

① ***Analyse***

Test bilatéral $n \geq 4$

On sait que :

$$\text{Prob}(\chi^2_{n-1,\alpha/2} \leq \frac{ns^2}{\sigma^2_0} \leq \chi^2_{n-1;1-\alpha/2}) = 1-\alpha$$

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si la valeur de $\frac{ns^2}{\sigma^2_0}$ appartient à *l'intervalle* :

$$[\ \chi^2_{n-1;\alpha/2};\ \chi^2_{n-1;1-\alpha/2}\ ]$$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon***.

IV - 2 TEST DE COMPARAISONS

Il s'agit de comparer deux variances selon la nature de la population.

IV - 2. 1 *Population quelconque*

✧ Supposons ***H_0 vraie*** $\sigma^2 = \sigma^2_0$

① ***Analyse***

On sait que les *intervalles de confiance, au niveau* α, sont donnés par :

***Pour** $\sigma_1^2 - \sigma_2^2$:*

$$S^2_1 - S^2_2 \pm t_{\alpha/2}\sqrt{\frac{\hat{\mu}^{(1)}_4 - \hat{\sigma}_1^4}{n_1} + \frac{\hat{\mu}^{(2)}_4 - \hat{\sigma}_2^4}{n_2}}$$

***Pour** $\sigma_1 - \sigma_2$*

$$S_1 - S_2 \pm t_{\alpha/2}\sqrt{\frac{\hat{\mu}^{(1)}_4 - \hat{\sigma}_1^4}{4n_1\hat{\sigma}^2_1} + \frac{\hat{\mu}^{(2)}_4 - \hat{\sigma}_2^4}{4n_2\hat{\sigma}_2}}$$

On utilise la *distribution normale* et on *remplace* les paramètres inconnus par leurs *estimations*.

② ***Décision***

✧ ***L' hypothèse H_0*** est *acceptée si*

$$\left| \frac{S^2_1 - S^2_2}{\sqrt{\frac{\hat{\mu}^{(1)}_4 - S_1^4}{n_1} + \frac{\hat{\mu}^{(2)}_4 - S_2^4}{n_2}}} \right| \leq t_{\alpha/2}$$

✧ *L'intervalle d'acceptation* correspondant à H_0 : $\sigma = \sigma_0$, est

$$\left| \frac{S_1 - S_2}{\sqrt{\frac{\hat{\mu}^{(1)}_4 - S_1^4}{4n_1S^2_1} + \frac{\hat{\mu}^{(2)}_4 - S_2^4}{4n_2S^2_2}}} \right| \leq t_{\alpha/2}$$

IV - 2. 2 ***Population approximativement normale***

✧ Supposons H_0 vraie $\sigma^2 = \sigma^2_0$

① ***Analyse***

En remplaçant $\sigma^2 - \sigma^2_0$ par 0, *les intervalles de confiance, au niveau* α, s'obtiennent par :

$$*\textbf{Pour } S^2_1 - S^2_2 : \qquad \pm t_{\alpha/2'} \sqrt{\frac{2S_1^4}{n_1} + \frac{2S_2^4}{n_2}}$$

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si la valeur de $S^2_1 - S^2_2$ appartient à l'intervalle de confiance précédent.

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***

✧ Pour ***l'hypothèse H_0*** relative à $\sigma = \sigma_0$:

* *on accepte H_0 lorsque $S_1 - S_2$ appartient à l'intervalle* :

$$\pm t_{\alpha/2'} \sqrt{\frac{S_1^2}{2n_1} + \frac{S_2^2}{2n_2}}$$

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***

IV - 2. 3 *Population normale*

✧ Supposons H_0 vraie $\sigma^2 = \sigma^2_0$

➢ ***Test bilatéral***

① ***Analyse***

Rappelons les résultats suivants :

❶ $\frac{ns^2}{\sigma^2} \rightarrow \chi^2_{n-1}$: *loi de Khi-deux à (n-1) degrés de liberté*

❷ $\frac{S^2_1}{S^2_2} = \frac{\frac{n_1 s^2_1}{n_1 - 1}}{\frac{n_2 s^2_2}{n_2 - 1}} = F_{n_1-1;n_2-1}$ avec $\sigma^2_1 = \sigma^2_2$

loi de Fisher-Snedecor à ($n_1 - 1$) et ($n_2 - 1$) degrés de liberté

❸ $\frac{n_1 s^2_1}{\sigma^2_1} - \frac{n_2 s^2_2}{\sigma^2_2} \rightarrow \chi^2_{n_1 - n_2}$ $n_1 > n_2$

Comme $\sigma^2_1 = \sigma^2_2 = \sigma^2$ et que cette variance est inconnue, elle ne peut servir pour le test. On utilise le rapport alors :

$$\frac{S^2_1}{S^2_2} = \frac{\frac{n_1 s^2_1}{n_1 - 1}}{\frac{n_2 s^2_2}{n_2 - 1}} = F_{n_1-1;n_2-1}$$

On doit s'attendre à avoir le *rapport égal à l'unité*.

On introduit *deux valeurs de la variable de Fisher* dépendant de α telles que :

$$\text{Prob}(F_{n_1-1;n_2-1} \leq F_{n_1-1;n_2-1,\alpha/2}) = \frac{\alpha}{2}$$

et $$\text{Prob}(F_{n_1-1;n_2-1} \leq F_{n_1-1;n_2-1,1-\alpha/2}) = 1 - \frac{\alpha}{2}$$

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité α,* si la valeur de $\frac{S^2_1}{S^2_2}$ appartient à l'intervalle de confiance précédent.

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***

✧ *L' hypothèse H_0 est également rejetée si ce rapport est trop petit ou trop grand*

Remarque

☞ $F_{n1-1;n2-1,\alpha/2}$ ne se trouve pas dans la table. On considère le rapport de la plus grande variance à la plus petite variance et on le compare à $F_{n1-1;n2-1,1-\alpha/2}$

✧ ***L'hypothèse H_0*** est *acceptée si* $\frac{S^2\max}{S^2\min} \leq F_{n1-1;n2-1,1-\alpha/2}$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon***

➢ ***Test unilatéral***

On peut être amené à tester une hypothèse H_0 : $\sigma^2_1 \geq \sigma^2_2$ par rapport à $\sigma^2_1 < \sigma^2_2$ au niveau α. Dans ce cas on considère le rapport inverse et *intuitivement* on rejette l'hypothèse quand le *rapport est trop grand.*

Cas de grands échantillons

① ***Analyse***

Ce cas peut être traité par l'introduction de la *variable z de Fisher* :

$$z = \frac{1}{2}\ln e\frac{S_1^2}{S_2^2}$$

et en considérant la variable $Z = \dfrac{z}{\sqrt{\frac{1}{2}(\frac{1}{n_1-1}+\frac{1}{n_2-1})}}$

comme une *variable normale centrée réduite* lorsque $\sigma^2_1 = \sigma^2_2$

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si la valeur observée appartient à l'intervalle $[-t_{\alpha/2};\ t_{\alpha/2}]$

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***.

Remarques

Echantillons appariés **:TEST DE MCNEMAR**

Pour tester deux moyennes de deux séries appariées, on considère la distribution d'échantillonnage de leur différence. *Deux cas à distinguer* :

a) σ connu

$m_d : m_1 - m_2$ la moyenne des différences suit *une loi normale* et on procède comme pour une seule moyenne tant au point de vue analyse que décision.

b) σ inconnu

① ***Analyse***

$m_d : m_1 - m_2$ la moyenne des différences suit une *loi de Student* de *degré (n – 1)* et on remplace σ par la quasi-variance S

On calcule alors $t_{observé} = \dfrac{m_d}{\dfrac{S_d}{\sqrt{n}}}$

et on le compare à la valeur trouvée $t(n-1)$ dans la table de Student

② ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité α,* si la *valeur numérique trouvée est inférieure ou égale à* $t(n-1)$

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon***.

IV - 3 ANALYSE DE LA VARIANCE : *ANOVA*

L'analyse de la variance, connue sous *l'acronyme ANOVA* permet de déterminer *l'influence d'un ou plusieurs facteurs sur la variable X.* Nous allons nous limiter aux cas les plus simples :

ANOVA a un seul facteur et ANOVA à deux facteurs.

IV- 3. 1 *ANOVA avec un seul facteur*

❑ Au lieu de deux populations, nous considérons ***k populations*** normales P ($k > 2$) de même variance σ^2.

❑ Pour ***k populations***, on affecte respectivement les notations précédentes par les indices *1, ..., k*:

$P_1 \Rightarrow m_1 ; \sigma^2 \qquad P_k \Rightarrow m_k ; \sigma^2$

❑ On prélève dans cette population ***k échantillons aléatoires*** de tailles respectives $n_{1,\dots,}\ n_k$ et on pose : $n = \sum_{i=1}^{k} n_i$

❑ On appelle

$x_{i,j}$ la $j^{ème}$ observation du $i^{ème}$ échantillon

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{i,j} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{x}_i$$

❑ On teste l'hypothèse nulle H_0 : $\quad m_1 = m_i = \dots = m_k$

① ***Analyse***

Le principe consiste à :

✧ obtenir, dans le cadre de ***l'hypothèse nulle H_0*** , *deux estimations indépendantes* de la *variance* commune σ^2

✧ *tester l'égalité de ces deux variances*

② ***Méthode***

On démontre facilement l'égalité suivante :

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{i,j} - \bar{x})^2 = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{i,j} - \bar{x}_i)^2$$

Pour simplifier, ces expressions seront désignées respectivement par :

Q : *la somme des carrés totale*

Q_1 : *la somme des carrés entre les suites d'observations*

Q_2 : *la somme des carrés dans les suites*

En divisant les différentes sommes par la variance commune et dans le cadre de l'hypothèse nulle $\boldsymbol{H_0}$, on a les distributions suivantes :

$$\frac{Q}{\sigma^2} \to \chi^2(n-1)\ ;\ \frac{Q_1}{\sigma^2} \to \chi^2(k-1)\ ;\ \frac{Q_2}{\sigma^2} \to \chi^2(n-k)$$

Les deux dernières expressions sont indépendantes.

En utilisant les *propriétés de l'espérance*, on obtient les *égalités suivantes* :

$$E\left(\frac{Q}{n-1}\right) = E\left(\frac{Q_1}{k-1}\right) = E\left(\frac{Q_2}{n-k}\right) = \sigma^2 \qquad \text{On pose } F = \frac{\dfrac{Q_1}{k-1}}{\dfrac{Q_2}{n-k}}.$$

F est *une variable de Fisher-Snedecor à (k – 1) et (n – k) degrés de liberté.* Pour un niveau de probabilité α, correspond une valeur F_α, lue dans la table, telle que Prob(F > F_α) = α

③ ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si la *valeur numérique de F est inférieure à* F_α

✧ ***L' hypothèse H_0*** est *rejetée* ***sinon***.

Remarques

☞ On présente généralement les résultats sous la forme d'un tableau appelé ***tableau d'analyse de la variance***

Somme des carrés	*Degré de liberté*	*Estimation de la variance*	
Q_1	*k - 1*	$\frac{Q_1}{k-1}$	Variance entre échantillons *Moyenne des carrés due au facteur*
Q_2	*n - k*	$\frac{Q_2}{n-k}$	Variance résiduelle *Moyenne des carrés due à l'erreur*
Q	*n - 1*		

☞ On calcule généralement Q et Q_1 ; Q_2 se *déduit par différence*

☞ Ces calculs se simplifient en utilisant :

S_i : la *somme* des observations de l'*échantillon i* tiré de la *population i*

S : *la somme de toutes les observations*

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x^2_{i,j} - \frac{S^2}{n} \; ; \; Q_1 = \sum_{i=1}^{k} \frac{S^2_i}{n_i} - \frac{S^2}{n}$$

IV - 3. 2 ***ANOVA avec deux facteurs***

❑ Il s'agit de tester les effets simultanés de deux facteurs, *A* et *B*

❑ Le facteur *A* avec *k variables* et le facteur *B* avec *r variables*

❑ Pour chaque combinaison A_iB_j, on considère un *seul résultat* $x_{i,j}$

① ***Analyse***

Le principe consiste à :

✧ tester l'égalité des moyennes relatives à chaque case du *tableau suivant* :

	B_1		B_j		Br	*moyenne*
A_1	x_{11}		x_{1j}		x_{1r}	$\bar{x}_{1.}$
A_i	x_{i1}		x_{ij}		x_{ir}	$\bar{x}_{i.}$
A_k	x_{k1}		x_{kj}		x_{kr}	$\bar{x}_{k.}$
moyenne	$\bar{x}_{.1}$		$\bar{x}_{.j}$		$\bar{x}_{.r}$	$\bar{x}$

② ***Méthode***

On pose : $\bar{x}_{i.} = \frac{1}{r}\sum_{j=1}^{r} x_{i,j}$ $\qquad \bar{x}_{.j} = \frac{1}{k}\sum_{j=1}^{k} x_{i,j}$ $\qquad \bar{x} = \frac{1}{rk}\sum_{i=1}^{k}\sum_{j=1}^{r} x_{i,j}$

On démontre l'égalité suivante :

$$\sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij}-\bar{x})^2 = r\sum_{i=1}^{k}(\bar{x}_{i.}-\bar{x})^2 + k\sum_{j=1}^{k}(\bar{x}_{.j}-\bar{x})^2 + \sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x})^2$$

Pour simplifier, les *expressions précédentes* seront désignées *respectivement* par Q , Q_A ; Q_B et Q_E.

En divisant les différentes sommes par la variance commune et dans le cadre de l'hypothèse nulle $\boldsymbol{H_0}$, on a les distributions suivantes :

$$\frac{Q}{\sigma^2} \to \chi^2(rk-1) \; ; \; \frac{Q_A}{\sigma^2} \to \chi^2(k-1) \; ; \; \frac{Q_B}{\sigma^2} \to \chi^2(r-1) \; ; \; \frac{Q_E}{\sigma^2} \to \chi^2(r-1)(k-1)$$

Les trois dernières expressions sont indépendantes.

✧ *Pour tester l'influence de A* et *l'influence de B* on définit :

$$F_A = \frac{\frac{Q_A}{k-1}}{\frac{Q_E}{(k-1)(r-1)}} \qquad \text{et} \qquad F_B = \frac{\frac{Q_B}{r-1}}{\frac{Q_E}{(k-1)(r-1)}}.$$

③ ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité α,* si la *valeur numérique de F est inférieure à F_α*

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***.

Remarques

☞ On présente les résultats sous la forme d'un tableau

Tableau d'analyse de la variance

Somme des carrés	*Degré de liberté*	*Estimation de la variance*	
Q_A	*k - 1*	$\frac{Q_A}{k-1}$	*Variance due au facteur A*
Q_B	*r -1*	$\frac{Q_B}{r-1}$	*Variance due au facteur B*
Q_E	*(r-1)(k-1)*	$\frac{Q_E}{(r-1)(k-1)}$	*Variance résiduelle*
Q	*rk - 1*		

☞ On *calcule généralement* Q ; Q_A et Q_B ; Q_E se *déduit par différence*

☞ Ces *calculs se simplifient* en utilisant la *somme des lignes*, la *somme des colonnes* et *la somme de toutes les observations* :

$$Q_1 = \frac{1}{k}\sum_{i=1}^{k} S^2_{i.} - \frac{S^2}{kr} \; ; \; Q_2 = \frac{1}{r}\sum_{j=1}^{r} S^2_{.j} - \frac{S^2}{kr} \; ; \; Q = \sum_{j=1}^{r}\sum_{i=1}^{k} x^2_{ij} - \frac{S^2}{kr}$$

☞ On peut considérer un *modèle plus général* que le précédent en faisant figurer les *interactions entre A et B*. Si le test d'interaction est significatif, on conclut que l'effet d'un facteur est fonction de l'autre.

☞ Des *logiciels* ont été développés pour effectuer les *calculs de l'analyse* de la variance : *SPSS* par exemple.

V - TEST SUR LE COEFFICIENT DE CORRELATION

❑ Le *coefficient de corrélation r*, calculé à partir d'un échantillon de taille *n*, est une variable aléatoire qui prend ses valeurs autour du coefficient de corrélation ρ de la population ***P***.

❑ On démontre que c'est *un estimateur convergent* et *efficace*.

❑ Selon la valeur du coefficient de corrélation ρ, on *distingue deux cas* :

a) $\underline{\rho = 0}$

① ***Analyse***

Dans ce cas la distribution de *r est symétrique*.

✧ ***L' hypothèse H_0*** : absence de corrélation linéaire , on teste $\rho = 0$ par rapport à H_1 ***où*** $\rho \neq 0$

On montre que la variable

$$\frac{r\sqrt{n-2}}{1-r^2} \rightarrow t_{n-2}$$

On suppose que l'échantillon a été prélevé d'une population normale à deux dimensions. On calcule la *valeur critique* de *r*

$$r_c = \frac{t_{n-2;\alpha/2}}{\sqrt{t^2_{n-2;\alpha/2} + (n-2)}}$$

② ***Décision***

On peut dresser le tableau suivant :

Hypothèses	Loi de Student t_{n-2}	Valeurs de r_c
H_0 : ρ = 0 H_1 : ρ ≠ 0	Rejet de H_0 si $t > t_{n-2,\alpha/2}$ ou $t < -t_{n-2,\alpha/2}$	Rejet de H_0 $r > r_c$ ou $r < r_c$

b) $\underline{\rho = \rho_0 \neq 0}$

① ***Analyse***

✧ ***L'hypothèse H_0*** où $\rho = \rho_0 \neq 0$

par rapport à ***H_1*** où $\rho < \rho_0$ *ou* $\rho > \rho_0$

On utilise la transformation de Fisher $Z' = \frac{1}{2}\ln\frac{1+r}{1-r}$

On démontre que ses *caractéristiques* sont données par :

$$E(z') = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \qquad \text{et} \quad \sigma(z') = \frac{1}{\sqrt{n-3}}$$

La variable centrée réduite est $T = \frac{z' - E(z')}{\sigma(z')} \rightarrow N(0,1)$

② ***Décision***

✧ ***L' hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si la valeur $T > t_{\alpha/2}$ ou $T < - t_{\alpha/2}$

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***.

Tableau-synthèse de comparaisons d'échantillons

Test : H_0 contre H_1	***Statistique***	***Loi de S*** sous H_0	***Règle de décision*** *valeur observée* $\geq C$
A) Un seul Echantillon $N(m ; \sigma)$ H_0 $m = m_0$ *connu* H_1 $m > m_0$ σ *inconnu*	$\dfrac{\overline{X} - m_0}{\dfrac{S}{\sqrt{n}}}$	T_{n-1}	$\dfrac{\overline{X} - m_0}{\dfrac{S}{\sqrt{n}}}$
H_0 $\sigma = \sigma_0$ H_1 $\sigma > \sigma_0$ σ_0 connu m_1 inconnu	S^2	$\dfrac{\sigma_0}{n-1}\chi^2(n-1)$	S^2
B) Des Echantillons ***Normaux*** H_0 $m_1 = m_2$ H_1 $m_1 > m_2$ $\sigma_1 = \sigma_2 = \sigma$ *inconnus*	$\dfrac{\overline{X} - \overline{Y}}{\hat{\sigma}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	$T_{n_1 + n_2 - 2}$	$\dfrac{\overline{X} - \overline{Y}}{\hat{\sigma}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$
H_0 $m_1 = m_2$ H_1 $m_1 > m_2$ σ_1, σ_2 *connus*	$\overline{X} - \overline{Y}$	$N(0, \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}})$	$\overline{X} - \overline{Y}$
H_0 $\sigma_1 = \sigma_2$ H_1 $\sigma_1 > \sigma_2$ m_1, m_2 *connues*	$\dfrac{S'^2_X}{S'^2_Y}$ S' *le meilleur*	$F(n_1, n_2)$	$\dfrac{S'^2_X}{S'^2_Y}$
H_0 $\sigma_1 = \sigma_2$ H_1 $\sigma_1 > \sigma_2$ m_1, m_2 *inconnues*	$\dfrac{S_X^2}{S_Y^2}$	$F(n_1 - 1, n_2 - 1)$	$\dfrac{S_X^2}{S_Y^2}$
Comparaisons de fréquences H_0 $p_1 = p_2$ H_1 $p_1 > p_2$	$\dfrac{f_1 - f_2}{\sqrt{f(1-f)(\dfrac{1}{n_1} + \dfrac{1}{n_2})}}$ où $f = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$	$N(0, 1)$	$\dfrac{f_1 - f_2}{\sqrt{f(1-f)(\dfrac{1}{n_1} + \dfrac{1}{n_2})}}$

Auto-évaluation ?

1- Formuler les hypothèses statistiques H_0 et H_1 correspondantes :

a) l'âge moyen des personnes de l'entreprise est supérieur à 30 ans.

b) la durée moyenne d'un pneu est de 50000 km, avec de nouvelles techniques de fabrications, on peut améliorer cette durée de vie.

c) le taux de cholestérol sérique est en moyenne de 2g/l.

d) le taux d'écoute d'un certain programme de télévision a été établi à 15%, à l'aide d'une meilleure présentation, on peut le rendre plus important.

e) Deux malades sur cinq sont influencés par la notoriété du médecin

2- Les risques d'erreur pour un test d'hypothèse sont

a) complémentaires

b) indépendants

3- Pour comparer une probabilité avec une valeur fixée, on utilise

a) le test de comparaisons de moyennes

b) le test de *Khi-deux*

4- On appelle seuil de confiance

a) la probabilité α

b) la probabilité $1-\alpha$

5- Combien y a-t-il de sortes de risques d'erreur pour un test d'hypothèse ?

6- Dans un test d'hypothèse, les conclusions possibles sont symétriques

a) Vrai b) Faux

7- Dans un test sur une proportion, l'hypothèse nulle s'écrit $H_0 (p = p_0)$ et l'hypothèse alternative H_1 s'écrit :

a) $p \neq p_0$ b) $p > p_0$ c) $p < p_0$

8- Dans un test sur une moyenne, l'hypothèse nulle H_0 est centrée sur :

a) $\bar{x}$ b) $m = m_0$

9- Si on rejette l'hypothèse nulle H_0 alors qu'elle est vraie, on commet une erreur de :

a) première espèce b) deuxième espèce

10- Pour calculer le risque de deuxième espèce, il ne faut aucune condition préalable :

a) Vrai b) Faux

11- Une réduction du risque de première espèce, réduit la zone de non-rejet de H_0

a) Vrai b) Faux

12- La variable centrée réduite, est utilisée si $np \geq 5$ et $nq \geq 5$

a) Vrai b) Faux

13- Quelles valeurs de s^2 d'un échantillon d'effectif n, issu d'une population normale, conduiraient au rejet de $H_0 : \sigma^2 = \sigma^2_0$ dans chacun des cas suivants :

$n = 21$ $\sigma^2_0 = 100$ $\alpha = 0{,}01$

$n = 15$ $\sigma^2_0 = 18$ $\alpha = 0{,}05$

14- Dans le cas de deux populations normales, tester à 5%

$H_0 : \sigma^2 = \sigma^2_1 = \sigma^2_2$ $n_1 = n_2 = 16$ $s^2_1 = 50$

$s^2_2 = 16$

15- Pour des populations presque normales construire un intervalle de sécurité pour $\sigma^2_1 - \sigma^2_2$.

Tester ensuite l'hypothèse $\sigma_1 = \sigma_2$ sur les données :

$n_1 = 180$; $n_2 = 160$ $s^2_1 = 47$ $s^2_2 = 36$

Exercices ?

Exercice 1

① Test bilatéral

On prélève, au hasard, un échantillon d'une population normale de moyenne μ, d'écart-type $\sigma = 8$ et de taille $n = 25$. On veut tester les hypothèses suivantes : $\mu = 52$ et $\mu \neq 52$.

1) Au seuil de signification 0,05, qu'elle est la règle de décision ?

2) On suppose que la moyenne de l'échantillon est $\bar{x} = 50$. Cette moyenne est-elle en contradiction avec l'hypothèse nulle H_0 ?

② Test unilatéral

On veut tester les hypothèses suivantes : $\mu = 52$ et $\mu > 52$.

Les hypothèses statistiques étant : $H_0 : \mu = 52$ et $H_1 : \mu > 52$.

Qu'elle est la probabilité de rejeter H_0 sachant que H_0 est vraie ? (au seuil de signification $\alpha = 0,05$)

Exercice 2

Un fabricant de chaussures désire mesurer l'influence d'une compagne publicitaire centrée sur Tunis. Un sondage effectué sur 100 magasins de la ville montre que sa marque est vendue par 60% des magasins interrogés. Après la compagne, un nouveau sondage réalisé sur 100 magasins, vendant sa marque, montre que 70% vendent sa propre marque.

1) Que peut-on conclure de cette progression ? Est-elle due en particulier à la compagne publicitaire ?

2) Même question pour $n = 200$ magasins.

Exercice 3

Une entreprise fabrique des ressorts dont la longueur X suit une loi normale $N(m = 4\ cm\ ;\ \sigma = 0,05\ cm)$. Pour vérifier la qualité de la production, on tire un échantillon de 25 ressorts. On considère que la qualité de la production est acceptable si la moyenne de l'échantillon $\bar{x}$ est comprise entre *3,985* cm et *4,015* cm

1) Calculer la probabilité pour que la production soit rejetée.
2) Quelle est la probabilité pour que l'écart-type de l'échantillon considéré soit inférieur à *0,04* cm ?

Exercice 4

Les observations sur deux prélèvements *N° 1* et *N°2* d'une caractéristique d'un produit *X* sont données par le tableau suivant :

N°1	53,1	52,85	54,30	53,75	53,55
N°2	53,2	52,55	52,31	52,23	

En supposant que X suit une distribution normale, tester s'il y a une différence significative entre les échantillons 1 et 2 si :

1) $\sigma_1 = \sigma_2 = 0{,}6$
2) $\sigma_1 = \sigma_2 = 1$
3) $\sigma_1 = 1$ et $\sigma_2 = 0{,}1$
4) $\sigma_1 = \sigma_2 = \sigma$ inconnu.

Exercice 5

On divise 300 sujets en deux groupes d'effectifs égaux, afin de tester deux médicaments contre l'hypertension. On observe 11 % d'amélioration pour le groupe 1 et 16% pour le groupe 2. Peut-on considérer que le médicament 1 est plus efficace que le médicament 2 ?

Exercice 6

Pour obtenir l'autorisation de mise sur le marché, un laboratoire pharmaceutique affirme que son médicament est efficace à 80%. Dans un centre hospitalier, on choisit au hasard, un échantillon de 200 individus et on leur administre le médicament. A l'issue d'une certaine période, on note que pour 150 individus, le traitement est efficace. Que peut-on conclure, au seuil de signification $\alpha = 0{,}05$ quant à l'affirmation du laboratoire ?

Exercice 7

Pour publier le *tome I* « ***Statistique & Probabilités*** », l'auteur exige de l'éditeur pour l'ensemble de la commande, au plus 2% de livres défectueux. L'auteur prélève au hasard 200 livres de la première livraison.

1) En utilisant un risque $\alpha = 0,05$ de rejeter à tort la livraison, quelle est la valeur critique de la proportion de livres défectueux qui ne doit pas être dépassée, pour accepter la production ?

2) Après avoir reçu l'ensemble de la commande, lors d'un contrôle général, l'auteur a observé 7 livres défectueux dans un échantillon de 200. La production peut-elle être considérée comme acceptable par l'auteur ?

3) L'auteur accepte la livraison des 200 premiers livres. Quelles sont les chances d'accepter 6% de livres défectueux, alors que l'exigence est de 2% dans 95% des cas ?

Exercice 8

Soit p la proportion d'objets défectueux sortant d'une machine. Si la machine est bien réglée $p = 0,10$; lorsqu'elle est déréglée $p = 0,20$. On veut choisir entre les deux décisions suivantes : ne rien faire ou réparer la machine.

1) Formuler les hypothèses statistiques H_0 et H_1 testées sur la base d'un échantillon de taille $n = 100$.

2) Calculer les probabilités à posteriori de H_0 et H_1

3) Qu'elle est la règle de décision optimale en appliquant la méthode de Bayes.

4) *Application numérique* :

$\mu = 0,6$ et $\frac{C_1}{C_2} = 2$ C_1 et C_2 désignant les coûts de réparation.

5- Si on observe 12 objets défectueux dans l'échantillon, quelle décision prendre ?

Exercice 9

On veut comparer la proportion p des naissances féminines à celle des naissances masculines grâce à un échantillon de 680 naissances.

1) On veut tester : $\begin{cases} H_0 : & p = 0,5 \\ H_1 : & p > 0,5 \end{cases}$

existe-t-il un test UMP pour tester H_0 contre H_1 (prendre $\alpha = 0,05$) ?

2) Même question si on veut tester : $\begin{cases} H_0 : & p = 0,5 \\ H_1 : & p \neq 0,5 \end{cases}$

3) Comment construire un 3 $^{\text{ième}}$ test pour un risque de $\alpha = 0,05$?

4) Sur 680 naissances, on a observé 306 filles à quelles décisions conduisent les tests précédents ?

Exercice 10

On suppose qu'on peut classer les lots de pièces achetés par une entreprise en deux catégories :

Catégorie A « production correcte » : la proportion des pièces défectueuses est de : 10% : $p = p_0 = 0{,}1$

Catégorie B « production défectueuse » : la proportion des pièces défectueuses est de : 20% : $p = p_1 = 0{,}2$

On effectue un sondage dans un lot, afin de choisir entre les deux hypothèses H_0 et H_1 :

Hypothèse H_0 : « il s'agit d'un lot provenant de la catégorie A »

Hypothèse H_1 : « il s'agit d'un lot de la catégorie B »

Sur un échantillon de n = 100 pièces prélevées au hasard, 14 sont défectueuses.

1) Définir le risque α de première espèce et le risque β de seconde espèce.

2) On fixe $\alpha = 0{,}05$ quel est l'intervalle d'acceptation de H_0 ? Quelle hypothèse doit-on admettre pour le lot considéré ?

3) Calculer le risque β de seconde espèce.

4) On fixe $\alpha = \beta = 0{,}05$. Combien faut-il prélever de pièces dans le lot?

Exercice 11

Avec l'engrais utilisé habituellement, le rendement d'un champ de blé (en quintaux par hectare) est une variable aléatoire suivant une loi normale d'espérance m = 50 et d'écart-type $\sigma = 12$. On admet qu'avec un nouvel engrais le rendement suit également une loi normale N(m ; σ). Pour savoir si le nouvel engrais donne plus de rendement que le premier, on l'essaie sur 80 parcelles et on observe :

$$\bar{x} = 52 \quad \text{et} \quad s^2 = \frac{\sum_{1}^{80} (x_i - \bar{x})^2}{n-1} = 180$$

1) Formaliser complètement le problème.

2) Quelle est la règle de décision pour un risque de 10% ?

3) Quelle est la décision à prendre au vu de l'échantillon ?

4) Quelle est la décision à prendre au vu de l'échantillon, si $m = 48$?

Exercice 12

Pour tester l'efficacité d'un nouveau médicament M préventif contre la grippe, on administre à un premier échantillon aléatoire X de taille $n_1 = 40$ le médicament et à un autre échantillon Y de taille $n_2 = 60$ un placebo P (une préparation ne contenant aucune substance active). A l'issue d'une certaine période, on note les résultats suivants :

	M	P
Pas de grippe	18	18
Grippe	22	42
Taille des échantillons	$n_1 = 40$	$n_2 = 60$

1) Formaliser le problème

2) Quelle est la règle de décision au seuil de 10%, de 3%.

3) On considère que la différence des fréquences observées est la même dans les deux échantillons de tailles respectives $n_1 = 40\,k$ et $n_2 = 60\,k$. A partir de quelle valeur de k, le médicament sera-t-il considéré comme efficace, au seuil de signification $\alpha = 0{,}05$?

Exercice 13

On se propose de comparer l'activité de 4 dispensaires de la région de Tunis. Pour chacun d'eux, on suppose que la distribution du nombre de consultants suit une loi normale, toutes de même variance. Les échantillons aléatoires, provenant de ces différents dispensaires, durant trois jours, sont donnés par le tableau suivant :

Dispensaire 1	60	60	39
Dispensaire 2	54	28	47
Dispensaire 3	48	52	53
Dispensaire 4	42	75	78

A partir de ces résultats, peut-on affirmer qu'il y a une différence significative de l'activité de ces dispensaires, au niveau $\alpha = 0{,}05$?

Corrigés

Auto-évaluation

1-

a $H_0: \mu = 30$ et $H_1: \mu > 30$

b $H_0: \mu = 50.000$ et $H_1: \mu > 50.000$

c $H_0: \mu = 2$ et $H_1: \mu \neq 2$

d $H_0: p = 0,15$ et $H_1: p > 0,15$

e $H_0: p = 0,4$ et $H_1: p \neq 0,4$

2- b

3- a; b

4- α

5- Deux sortes de risques : risque de première espèce α ; risque de seconde espèce β

6- b (si on rejette l'hypothèse nulle, on accepte un risque d'erreur α). Mais on ne peut pas accepter l'hypothèse.

7- a ; b ; c

8- b

9- a

10- b (il faut spécifier une valeur particulière du paramètre dans l'hypothèse H_0 que l'on suppose vraie).

11- b

12- a

13- *Population normale* :

H_0: $\sigma^2 = \sigma^2_0$

① $n = 21$ $\quad \sigma^2_0 = 100 \quad$ $\alpha = 0,01$

$$\frac{n s^2}{\sigma^2} \rightarrow \chi^2_{n-1} \quad \text{et} \quad \frac{21 s^2}{\sigma^2} \rightarrow \chi^2_{20}$$

$$\text{Prob}\left(a < \frac{21 s^2}{\sigma^2} < b\right) = 0,99 \Rightarrow \text{Prob}\left(7,43 < \frac{21 s^2}{\sigma^2} < 40\right) = 0,99$$

$$\Rightarrow \text{Prob}\left(0,525\, s^2 < \sigma^2 < 2,826\, s^2\right) = 0,99$$

Pour que = soit rejetée comme hypothèse, il faudrait :

$0,525\, s^2 > 100$ ou $2,826\, s^2 < 100 \Rightarrow s > 13,8$ ou $s < 5,95$

② $n = 15 \quad \sigma^2_0 = 18 \quad \alpha = 0,05$

$$\frac{15 s^2}{\sigma^2} \rightarrow \chi^2_{14} \Rightarrow \text{Prob}\left(5,63 < \frac{15 s^2}{\sigma^2} < 26,12\right) = 0,95$$

$$\Rightarrow \text{Prob}\left(0,574\, s^2 < \sigma^2 < 2,664\, s^2\right) = 0,95$$

Pour que $\sigma^2 = 100$ soit rejetée comme hypothèse, il faudrait :

$0,574\, s^2 > 18$ ou $2,664\, s^2 < 18 \Rightarrow s > 5,6$ ou $s < 2,6$

14- On teste à 5% c'est à dire, il faut chercher deux valeurs F' *et* F'' telles que :

$$\text{Prob}\left(F_{n1-1,\, n2-1} < F'\right) = 0,025 \; ; \text{Prob}\left(F_{n1-1,\, n2-1} < F''\right) = 0,975$$

$H_0: \sigma^2 = \sigma^2_1 = \sigma^2_2 \quad n_1 = n_2 = 16 \quad s^2_1 = 50 \quad s^2_2 = 16$

$F_{15,15} = 3{,}125 \frac{\sigma^2_2}{\sigma^2_1}$. Par lecture dans la *table de Fisher-Snedecor,* on trouve :

$$F' = 0{,}35 \quad et \quad F'' = 2{,}86 \Rightarrow \text{Prob}\,(1{,}09 < \frac{\sigma^2_1}{\sigma^2_2} < 8{,}93) = 0{,}95$$

Pour que $\sigma^2_1 = \sigma^2_2$, il faudrait que 1 appartienne à l'intervalle [*1,09 ; 8,93*]

Or, il n'en fait pas partie $\Rightarrow$ σ^2_1 et σ^2_2 auront 95% de chance d'être différents.

15- *Populations presque normales*

Construisons un intervalle de confiance (de sécurité) pour $\sigma^2_1 - \sigma^2_2$. *à 95%.*

On sait que : $s^2 \rightarrow N\,(\sigma^2, s^2\sqrt{\frac{2}{n}})$

$n_1 = 180 \qquad n_2 = 160 \qquad s^2_1 = 47 \qquad s^2_2 = 36$

$T_1 = \frac{47 - \sigma^2_1}{4{,}95} \rightarrow N\,(0,1) \qquad$ et $\qquad T_2 = \frac{36 - \sigma^2_2}{4{,}025} \rightarrow N\,(0,1)$

Prob (37,298 < σ^2_1 < 56,702) = 0,95 et Prob (28,111 < σ^2_2 <43,889) = 0,95

Prob(9,187 < $\sigma^2_1 - \sigma^2_2$. < *12,813) = 0,95*

Pour que $\sigma^2_1 = \sigma^2_2$, il faudrait que 0 appartienne à l'intervalle

[*9,187; 12,813*].

Or, il n'en fait pas partie $\Rightarrow$ σ^2_1 et σ^2_2 auront 95% de chance d'être différents.

Exercices

Exercice 1

① ***Test bilatéral***

Hypothèses statistiques : $H_0 : \mu = 52$ et $H_1 : \mu \neq 52$

Seuil de signification : $\alpha = 0,05$

Echantillon aléatoire d'une population normale d'écart-type $\sigma = 8$

La statistique est la variable centrée réduite :

$$T = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad \text{avec } \mu = 52$$

L'intervalle de $\bar{x}$ est donné par $\bar{x} \in [\mu \pm t_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$

En remplaçant $t_{0,025} = \pm 1,96$ [valeur donnée par la *loi N(0, 1)*], on obtient $t \in$ [48,864 ; 55,136] au seuil $\alpha = 0,05$

Règle de décision :

On rejette H_0 si $\bar{x} < 8,864$ ou $\bar{x} > 8,864$ et on ne la rejette pas ailleurs. Cette hypothèse n'est pas en contradiction avec l'hypothèse nulle H_0

② ***Test unilatéral***

La valeur critique 54,632 correspond à $t_{0,05} = 1,645$

Règle de décision

On rejette H_0 si $\bar{x} > 54,632$ et on ne la rejette pas ailleurs

Exercice 2

① La différence est-elle significative ?

$f_1 = 0,6$ et $f_2 = 0,7$ comme n > 30 $\qquad f_1 - f_2 \rightarrow N(p, \sqrt{\frac{pq}{n}})$

p étant la proportion réelle.

Construction du test :

Soit p_1 la proportion du test dans l'ensemble des magasins avant la campagne et p_2 après la campagne : $p_1 \neq p_2$

$H_0 : \quad p_1 = p_2 = p \qquad$ alors $\qquad f_p = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$

$$T = \frac{f_1 - f_2}{\sqrt{f_p(1-f_p)(\frac{1}{n_1}+\frac{1}{n_2})}} \rightarrow N(0,\ 1)$$

Seuil de confiance: $\alpha = 0,05$ on a $| f_1 - f_2 | < 1,96\sqrt{0,65\times0,35\times\frac{1}{50}} = 0,132$

Or la différence des fréquences constatées est de 0,1, elle est inférieure à 0,13 on ne peut l'attribuer avec certitude à l'influence de la publicité mais plutôt aux variations d'échantillonnage. Les résultats de l'enquête ne peuvent affirmer que l'influence de la publicité est nulle.

② En augmentant la taille de l'échantillon

$$| f_1 - f_2 | < 1,96\sqrt{0,65\times0,35\times\frac{1}{100}} = 0,0935$$

La différence des fréquences constatées 0,1, est supérieure à 0,0935 d'où l'influence de la publicité.

Exercice 3

① *Probabilité de rejet* :

On calcule la probabilité d'acceptation : Prob(3,985 < $\bar{x}$ < 4,015). En utilisant la variable centrée réduite, on obtient :

Prob(-1,5 < T < 1,5) = 2G(1,5) – 1 = 0,8664

d'où la *probabilité de rejet : 0,1336*

② *Prob(s < 0,04)*

On sait que $\text{Prob}[(\frac{(n-1)s^2}{\sigma^2} < \frac{16\times10^{-4}\times24}{25\times10^{-4}}]$ d'où Prob(s^2) = 16 10^{-4} et en lisant dans la *table de la loi de khi-deux*, pour un degré de *liberté égal à 24*, on obtient : $\text{Prob}(\chi^2_{24} < 15,36) \# 0,1$

Exercice 4

X_1	53,1	52,85	54,30	53,75	53,55	267,55
X_2	53,2	52,55	52,31	52,23		210,29
X^2_1	2819,6	2793,1225	2948,49	2889,0625	2867,6025	14317,8775
X^2_2	2830,24	2761,5025	2736,3361	2727,9729		11056,0515

La taille des échantillons étant respectivement $n_i = 4$ et $n_2 = 5$, on calcule les moyennes puis les écarts-types respectifs, on trouve :

$$\bar{x}_1 = \frac{267,55}{5} = 53,51 \; ; \; \bar{x}_2 = \frac{210,29}{4} = 52,5725$$

$$s^2_1 = 0,2554 \; ; s^2_2 = 0,14512$$

① $\sigma_i = \sigma_2 = \sigma = 0,6$

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sigma\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}} \rightarrow N(0, 1) \text{ au seuil } \alpha = 0,05 \Rightarrow t = 1,96$$

Prob [$0,9375 - 0,7889 \leq m_1 - m_2 \leq 0,9375 + 0,7889$]= 0,95

L'intervalle est donc [$0,1486 \leq m_1 - m_2 \leq 1,7264$]

Comme 0 n'appartient pas à cet intervalle, m_1 *et* m_2 auront 95 % de chances d'être différents.

② $\sigma_1 = \sigma_2 = \sigma = 1$

En utilisant la variable centrée réduite T et en effectuant les calculs comme précédemment, on obtient l'intervalle :

$$[-0,3777 \leq m_1 - m_2 \leq 2,2523]$$

Comme 0 appartient à cet intervalle, les deux procédés pourront être les mêmes.

③ $\sigma_1 = 1 \qquad \sigma_2 = 0,1$

Utilisant le même procédé, on trouve l'intervalle:

$$[0,3394 \leq m_1 - m_2 \leq 1,8411]$$

Comme 0 n'appartient pas à cet intervalle, m_1 *et* m_2 auront 95 % de chances d'être différents.

④ $\sigma_1 = \sigma_2 = \sigma$ *inconnu*

On utilise *la loi de Student*. On cherche sur la table la valeur de t pour 0,975 et ddl = 7 soit $t = 2,365$

$$\frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{S\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}} \rightarrow t_{n_1+n_2-2}$$

avec $S^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \Rightarrow S = 0,1628$

En remplaçant, on obtient l'intervalle : [$0,6847 \leq m_1 - m_2 \leq 1,1903$]

Comme 0 n'appartient pas à cet intervalle, m_1 *et* m_2 auront 95 % de chances d'être différents.

Exercice 5

Supposons l'effet des deux médicaments identiques, alors les fréquences observées ne doivent différer de la valeur théorique commune p que par des fluctuations d'échantillonnage.

Soit f la fréquence de guérison du médicament correspondant à 150 individus. On sait que $E(f) = p$ et $\text{Var}(f) = \frac{p(1-p)}{n} = \frac{p(1-p)}{150}$.

En appliquant le théorème central limite, on obtient :

$$\frac{f-p}{\sqrt{p(1-p)}}\sqrt{150} \rightarrow N(0,\ 1)$$

En estimant la valeur commune par

$$p = \frac{11+16}{2\times100} = 0{,}135 \quad \text{et V(f)} = 0{,}779\ 10^{-3} \text{ et } \sigma = 2{,}79\ 10^{-2}$$

Pour le premier groupe : $t_1 = \frac{f_1 - p}{\sigma} \rightarrow N(0,1)$

Pour le second groupe : $t_2 = \frac{f_2 - p}{\sigma} \rightarrow N(0,1)$

La différence t_1 - t_2 est donc une variable aléatoire :

$$\frac{t_1 - t_2}{\sqrt{2}} = \frac{f_2 - f_2}{\sigma\sqrt{2}} \quad \Rightarrow \text{ t \# 1,3}$$

$\Rightarrow$ les deux fréquences ne diffèrent pas d'une façon significative.

Exercice 6

Il s'agit de tester $H_0 \quad p = p_0$ au seuil $\alpha = 0{,}05 \Rightarrow \quad t = 1{,}96$

Règle de décision :

On rejette l'affirmation si $|T| < 1{,}96$

$$T = \frac{0{,}75 - 0{,}8}{\sqrt{\frac{0{,}8 \times 0{,}2}{200}}} = -1{,}767$$

Comme cette valeur n'est pas dans l'intervalle d'acceptation, on rejette l'hypothèse de la firme.

Exercice 7

① Il s'agit de tester H_0 $\quad p = p_0 = 0,02$ $\quad$ au seuil $\alpha = 0,05$

$$T = \frac{\hat{p} - 0,02}{\sqrt{\frac{0,98 \times 0,02}{200}}} = 1,645$$

d'où la proportion de livres à ne pas dépasser $\quad \hat{p} = 0,0363$

② La valeur calculée $\hat{p}' = 0,035$ est inférieure à 0,0363, la production est acceptable.

③ Il s'agit de déterminer la probabilité d'accepter H_0 alors qu'en réalité, la proportion de défectueux $H_1 : \hat{p} = 0,06$. Cette probabilité correspond au risque de 2^{ième} espèce : $\quad \beta = \text{Prob}(\hat{p} < 0,03 / p = 0,06 \text{ vraie})$

En utilisant la loi normale centrée réduite :

$$T = \frac{0,03 - 0,06}{\sqrt{\frac{0,94 \times 0,06}{200}}} = -1,786$$

Par lecture de la table, on a :

$\text{Prob}(0,03 \leq \hat{p} \leq 0,06) = 0,4633 \Rightarrow \beta = 0,5 - 0,4633 = 0,0367 \ \# \ 0,04$

Dans 4 % des cas, l'auteur acceptera des lots dont la proportion réelle de défectueux est de 0,06.

Exercice 8

① *Hypothèses statistiques possibles*

$$\begin{cases} H_0 & p_0 = 0,10 \\ H_1 & p_1 = 0,2 \end{cases}$$

Il s'agit de choisir entre deux décisions :

$$\begin{cases} D_0 & \text{ne pas réparer la machine} \\ D_1 & \text{réparer} \end{cases}$$

Tableau des décisions

Décisions / Hypothèses	D_0	D_1	probabilité à priori	probabilité à posteriori
H_0	C_{00}	C_{01}	μ	π_0
H_1	C_{10}	C_{11}	$1 - \mu$	π_1

② *Calcul des probabilités à posteriori* :

Soit X la variable aléatoire prenant les valeurs :

$$\begin{cases} 1 & \text{si l'objet est défectueux} \\ 0 & \text{sinon} \end{cases}$$

Sous les hypothèses :

$$H_0 \qquad X \to B(1;\ 0{,}1)$$
$$H_1 \qquad X \to B(1;\ 0{,}2)$$

Les vraisemblances respectives de l'échantillon :

$$L_0(x_1,\ldots x_n, \theta) = p_0^{\sum_i^n x_i}(1-p_0)^{n-\sum_i^n x_i}$$

$$L_1(x_1,\ldots x_n, \theta) = p_1^{\sum_i^n x_i}(1-p_1)^{n-\sum_i^n x_i}$$

$$\pi_0 = \text{Prob}(H_0 / x_1, \ldots, x_n) = \frac{\text{Prob}(H_0)\times L_0}{\text{Prob}(H_0)\times L_0 + \text{Prob}(H_1)\times L_1}$$

et $$\pi_1 = \text{Prob}(H_1 / x_1, \ldots, x_n) = \frac{\text{Prob}(H_1)\times L_1}{\text{Prob}(H_0)\times L_0 + \text{Prob}(H_1)\times L_1}$$

$$\text{Prob}(H_0) = \mu \qquad et \qquad \text{Prob}(H_1) = 1 - \mu$$

Le coût des décisions est une variable aléatoire prenant les valeurs :

$$\begin{cases} 0 \text{ avec la probabilité } p_0 \\ C_1 \text{ avec la probabilité } p_1 \end{cases}$$

③ *La décision optimale* est celle qui minimise l'espérance mathématique du coût de chaque décision, soit :

$E(C_{D0})=C_1\pi_1$ et $E(C_{D1})=C_0\pi_0$

La règle de décision est celle qui fait correspondre l'espérance de coût la plus faible. Soit :

$$\frac{\pi_1 C_1}{\pi_0 C_0} = \frac{(1-\mu)\, L_1 C_1}{\mu\, L_0 C_0}$$

En calculant $\frac{L_1}{L_0}$ en passant aux logarithmes et en posant

$$k = \frac{\ln(\frac{C_0}{C_1} \times \frac{\mu}{(1-\mu)}) - n \ln(\frac{1-p_1}{1-p_0})}{\ln(\frac{1-p_0}{1-p_1} \times \frac{p_1}{p_0})}$$

On est conduit aux décisions suivantes :

$$\sum_{i=1}^{n} x_i < k \Rightarrow D_0 \quad ; \qquad \sum_{i=1}^{n} x_i \geq k \Rightarrow D_1$$

④ *Application numérique* :

Les calculs donnent k = 15,9 # 16. La règle de décision optimale est

$$\sum_{i=1}^{n} x_i \leq 15 \Rightarrow D_0 \quad ; \qquad \sum_{i=1}^{n} x_i \geq 16 \Rightarrow D_1$$

⑤ Si on a observé 12 objets défectueux sur 100, on décide D_0 : ne pas réparer la machine.

Exercice 9

① ***Test UMP***

Soit X la variable prenant les valeurs

$\begin{cases} 1 & \text{s'il y a naissance d'une fille} \\ 0 & \text{sinon} \end{cases}$

n = 680 fixons une valeur p_0 de p dans l'hypothèse H_1 le test UMP (le plus puissant) de seuil α, est défini par sa région critique : région de rejet de l'hypothèse de base H_0. On est conduit comme pour l'exercice précédent à utiliser les vraisemblances de l'échantillon et à calculer

$$\frac{L_1}{L_0} = (\frac{p_0}{p_1})^{\sum_i^n x_i} (\frac{1-p_0}{1-p_1})^{n-\sum_i^n x_i} \leq k$$

La région critique est définie par :

$$\frac{\sum_1^n x_i}{n} \geq C \qquad \alpha \text{ étant fixé : } \alpha = \text{Prob}(\frac{\sum_1^n x_i}{n} \geq C / p = 0{,}5)$$

Comme n est grand, on utilise la loi normale centrée réduite

$$\alpha = \text{Prob}(T \geq \frac{C - 0{,}5}{\sqrt{\frac{(0{,}5)^2}{680}}})$$

avec *T suivant N(0, 1)*. Les calculs donnent *C # 0,532*. La région critique est indépendante de la valeur de p_1 donc le test est UMP.

② Il s'agit de tester :

$$\begin{cases} H_0 & p_0 = 0{,}5 \\ H_1 & p_1 = 0{,}5 \end{cases}$$

La région critique est de la forme :

$$\frac{\sum_1^n x_i}{n} \geq C \text{ si } p_1 > 0{,}5 \; ; \; \frac{\sum_1^n x_i}{n} \geq C \text{ si } p_1 < 0{,}5$$

La région critique est dépendante de la valeur de p_1 donc le test n'est pas UMP.

③ Prenons comme test la réunion des deux régions précédentes. La région critique est définie par :

$$\frac{\sum_1^n x_i}{n} \geq C \text{ telle que : } \text{Prob}(\frac{\sum_1^n x_i}{n} \geq C / p = 0{,}5) = 0{,}025$$

$$\alpha = 0{,}025 \Rightarrow \text{ C \# 0,534}$$

Par analogie: $\text{Prob}(\frac{\sum_1^n x_i}{n} \leq C' / p = 0{,}5) = 0{,}025$, et par symétrie : C # 0,466

Le test proposé aura pour région critique l'une des régions précédentes.

④ On calcule $p = \frac{306}{680} = 0{,}45$

le test choisi en ① conduit à rejeter l'égalité des proportions des naissances et le test ② conduit à accepter l'égalité de ces proportions.

Exercice 10

Il s'agit de tester :

$$\begin{cases} H_0 & p = 0{,}1 \\ H_1 & p_1 = 0{,}20 \end{cases}$$

① *Risque* α = Prob(décider (H_1 / H_0)

X « *nombre de pièces défectueuses dans un échantillon de taille n* ». On sait que $f_n = \frac{X}{n} \to N(m, \sigma)$ lorsque n grand. Si l'on désigne π par la valeur critique :

Règle de décision :

$f_n \leq \pi$ *on décide* H_0

$f_n > \pi$ *on décide* H_1

d'où $\alpha = \text{Prob}(\text{décider } (H_1/H_0) = \text{Prob}(f_n \geq \pi / p = p_0)$

$\alpha = \text{Prob}(X \geq n\pi / p = p_0)$ de même $\beta = \text{Prob}(X \leq n\pi / p = p_1)$

② *Pour* $\alpha = 0{,}05$

En remplaçant par les valeurs numériques, la variable centrée réduite $T = \frac{\pi - p}{\sqrt{\frac{pq}{n}}}$ a pour valeur $t_\alpha = 1{,}646$. Comme $\pi = 14{,}95 \# 15$, *le lot est alors accepté* puisque $X = 14 < 15$.

③ *Risque* β

En remplaçant p par p_1, on obtient pour valeur de la variable $t = -1{,}26$ ce qui correspond à $\beta \# 0{,}1$

④ *Calcul de n*

$$\alpha = \beta = 0{,}05 \quad p_0 + t_\alpha \sqrt{\frac{p_0 q_0}{n}} = p_1 + t_\alpha \sqrt{\frac{p_1 q_1}{n}} \text{ d'où } n \# 133$$

Exercice 11

① *Formulation* : il s'agit de tester, avec *m inconnu*,

$$\begin{cases} H_0 & \sigma = 12 \\ H_1 & \sigma \neq 12 \end{cases}$$

La statistique utilisée est

$$S^2 = \frac{\sum_1^{80} (x_i - \bar{x})^2}{n-1} \text{ sa loi est telle que : } \frac{(n-1)s^2}{\sigma^2} \to \chi^2_{n-1}$$

La région critique de ce test est définie comme suit :

$$R = R_1 \cup R_2$$

avec $R_1 = \{ (x_1, \ldots, x_n) / s^2 \geq C \}$ avec $\text{Prob}(R_1 / H_0) = \frac{\alpha}{2}$

et $R_2 = \{ (x_1, ..., x_n) / s^2 \leq C' \}$ avec $\text{Prob}(R_2 / H_0) = \frac{\alpha}{2}$

② *Règle de décision*

Les valeurs de C et C' sont déterminées comme suit :

$\text{Prob}(S^2 \geq C / H_0) = \frac{\alpha}{2} = 0{,}05$ et $\text{Prob}(S^2 \leq C' / H_0) = \frac{\alpha}{2} = 0{,}05$

Puisque n est très grand, on peut considérer que : $\sqrt{2\chi^2} - \sqrt{2\nu - 1} \rightarrow N(0,1)$

En utilisant la loi normale et avec $\sigma = \sigma_0$

$$T = \sqrt{\frac{(n-1)\,s^2}{\sigma^2}} - \sqrt{2n - 3} \rightarrow N(0,1) \text{ on obtient :}$$

$$\sqrt{\frac{(n-1)\,C}{\sigma^2_0}} - \sqrt{2n - 3} = 1{,}65 \Rightarrow C = \frac{(1{,}65 + \sqrt{2n-3})^2}{2(n-1)} \sigma_0 = 183{,}25$$

de la même manière, on obtient *C' = 107,88*

③ Comme

$$R = \{ (x_1, ..., x_n) / s^2 \geq 183{,}25 \text{ ou } s^2 \leq 107{,}88 \}$$

la valeur donnée de $s^2 = 180$ appartient à l'intervalle de R, on décide d'accepter l'hypothèse $\sigma = 12$

④ *On teste* :

$$\begin{cases} H_0 & m = 48 \\ H_1 & m > 48 \text{ au seuil } \alpha = 0{,}1 \text{ avec } \sigma \text{ inconnu} \end{cases}$$

La statistique utilisée :

$t = \frac{\bar{x} - 48}{\frac{S}{\sqrt{n}}}$ avec t suivant la *loi de Student* de *(n – 1)* degrés de liberté.

Comme *n* est grand, t suit alors une loi normale centrée réduite et la région critique est définie par :

$$R = \{ (x_1, ..., x_n) / t \geq C \} \text{ avec } \text{Prob}(T \geq C) = \alpha = 0{,}1 \text{ sous } H_0$$

Soit C = 1,29 $t = 2{,}67$

On décide d'accepter l'hypothèse $m > 48$: *l'engrais est plus efficace.*

Exercice 12

① *Formulation* :

Soit X la variable aléatoire prenant les valeurs :

$$\begin{cases} 1 & \text{si la personne a la grippe avec } M \\ 0 & \text{sinon} \end{cases}$$

X suit une loi de *Bernoulli B(1, p_0)* où p_0 est la probabilité d'attraper la grippe quand on prend le médicament M.

Soit Y la variable aléatoire prenant les valeurs :

$$\begin{cases} 1 & \text{si la personne a la grippe sans } M \\ 0 & \text{sinon} \end{cases}$$

Y suit une loi de *Bernoulli B(1, p_1)* où p_1 est la probabilité d'attraper la grippe quand on ne prend pas le médicament M.

Il s'agit de tester :

$$\begin{cases} H_0 & p_0 = p_1 & M \text{ est inefficace} \\ H_1 & p_0 < p_1 & M \text{ est efficace} \end{cases}$$

② *Règle de décision* :

La région critique est telle que :

$$T = \frac{f_1 - f_2}{\sqrt{f\,(1-f\,)(\frac{1}{n_1} + \frac{1}{n_2})}} \leq C \text{ sous } H_0$$

$$f_1 = \frac{\sum_{i=1}^{n} x_i}{n_1} \qquad f_2 = \frac{\sum_{i=1}^{n} x_i}{n_2} \qquad \text{et} \qquad f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

La valeur de C est déterminée par $\text{Prob}(R / p_0 = p_1) = \alpha$

Soit : $\text{Prob}\,(T \leq C) = \alpha$

③ *Décision au seuil de 10 % et de 3 %*

En effectuant les calculs numériques, on obtient :

- $\alpha = 0{,}1 \Rightarrow C = -1{,}29$ et $t = -1{,}54$

Cette valeur est inférieure à (–1,29) $\Rightarrow$ M peut être considéré comme *efficace.*

- $\alpha = 0{,}03 \Rightarrow C = -1{,}88$ et $t = -1{,}53$

Cette valeur est supérieure à (–1,88) $\Rightarrow$ M peut être considéré comme *non efficace.*

③ Pour *décider* que M est efficace en tenant compte des *deux observations* à la fois, au seuil de 3 %, il faut que la valeur de t appartienne à la région critique soit :

$$t = \frac{f_1 - f_2}{\sqrt{f\,(1-f\,)(\frac{1}{n_1 k} + \frac{1}{n_2 k})}} \leq -1{,}88$$

avec : $-1{,}53\sqrt{k} = -1{,}88 \Rightarrow k \geq 1{,}51$

Pour un risque de 3 % la différence de fréquence sera significative dès que

$$n_1 + n_2 \geq 1{,}51 \quad \text{ou} \quad 3n_1 = 2\,n_2$$

Exercice 13

On teste l'hypothèse $H_0 : m_1 = m_2 = m_3 = m_4$

Tableau de l'ANOVA

				somme	Carré des sommes	moyenne
1	60	60	39	159	25281	53
2	54	28	47	129	16641	43
3	48	52	53	153	23409	51
4	42	75	78	195	38025	65
Σ				636	103356	53

$$Q_1 = \frac{103356}{3} - \frac{636^2}{12} = 744 \quad \Rightarrow \frac{Q_1}{3} = 248$$

$$Q_2 = 35920 - \frac{103356}{3} = 1468 \quad \Rightarrow \frac{Q_2}{8} = 183{,}5 \quad \Rightarrow F_0 = \frac{248}{183{,}5} = 1{,}35$$

	Somme des carrés	*Degré de liberté*	*Moyenne des carrés*	*Valeur observée pour F*
Facteur	744	3	248	1,35
Erreur	1468	8	183,5	
Total		11		

Comme F_0 = 1,35 est inférieure à $F_{0,05}$ = 4,07, on ne peut pas rejeter l'hypothèse H_0 . Au niveau de probabilité 0,05, on ne peut pas affirmer qu'il y a une différence significative dans l'activité de ces dispensaires.

Chapitre 6

TESTS NON PARAMETRIQUES

☛ *Objectifs*

- ⇨ Introduire le test de χ^2
- ⇨ étudier les applications de ce test
- ⇨ présenter les autres tests non paramétriques

☛ *Mots clés*

Test statistique du χ^2 – ajustement – fonction de répartition – conditions d'application du χ^2 – correction de Yates – test d'homogénéité – test d'indépendance – coefficient de contingence – autres tests.

I ***Test du χ^2***
II ***Test des signes***
III ***Test des rangs***
IV ***Test des séquences***

AUTO- EVALUATION & EXERCICES CORRIGES

Dans les chapitres précédents, nous avons passé en revue comment on pouvait tester des hypothèses relatives à des paramètres inconnus d'une population ***P***. La *distribution* de ***P*** a été supposée *connue* ou bien on a considéré que les conditions d'application du *théorème central-limite* sont vérifiées.

A défaut *d'appliquer ces méthodes*, d'autres tests dits *non-paramétriques* peuvent être utilisés c'est l'objet de ce chapitre.

Il s'agit de tester des hypothèses relatives à la *forme analytique* des distributions. On développera surtout le *test de* χ^2 et on présentera *d'autres tests* considérés comme usuels.

I - TEST DU KHI-DEUX

Dans le manuel « **Statistique & Probabilités** » il a été mentionné que si l'on ajuste une distribution théorique, on est amené à juger de la validité de cet ajustement à l'aide du ***test du khi-deux*** (χ^2) ou ***test de Pearson***. Le test du ***khi-deux*** (χ^2) permet de comparer des distributions entre elles et de vérifier si une *distribution expérimentale* suit une *distribution théorique* usuelle :$B(1 ; p)$; $B(n ; p)$; $P(\lambda)$; $N(m ; \sigma)$...

I - 1 TEST D'AJUSTEMENT

I - 1. 1 Variable χ^2

❶ ***Analyse***

Considérons une *variable aléatoire* X et un *échantillon aléatoire* de taille n : $(x_1,\ldots, x_n)$ et supposons que sa fonction de répartition $F(x)$ est celle d'une loi usuelle, complètement déterminée par ses paramètres. Il s'agit de tester l'hypothèse H_0 : *« X est une variable aléatoire ayant la fonction de répartition F(x) » c'est à dire de décider si les écarts entre la théorie* et *l'observation peuvent être imputés* au hasard.

❷ ***Construction du*** χ^2

On divise le domaine de variation de X en un nombre fini d'intervalles, mutuellement exclusifs et de longueurs quelconques de la forme:

$[a_0 ; a_1[;\ldots ; [a_{r-1} ; a_r[$. On désigne ces intervalles par $I_1 ;\ldots ; I_r$

La probabilité pour qu'une observation appartienne à la classe $[a_{i-1} ; a_i[$ est donnée par la fonction de répartition définie par :

$p_i = Prob\ \{x \in [a_{i-1}\ ;\ a_i[$ tel que $F(x)\}$ où $p_i = \int_{a_{i-1}}^{a_i} f(x)dx$

Soit n_i $(i = 1, 2, \ldots, r)$ le nombre de valeurs de l'échantillon appartenant à I_i, on a :

$$\sum_1^r p_i = 1 \text{ et } \sum_1^r n_i = n$$

Classes	Effectifs	Valeurs de X	Fréquences	Probabilités
$[a_0\ ;\ a_1[$	n_1	x_1	$\frac{n_1}{n}$	p_1
$[a_{i-1}\ ;\ a_i[$	n_i	x_i	$\frac{n_i}{n}$	p_i
$[a_{r-1}\ ;a_r[$	n_r	x_r	$\frac{n_r}{n}$	p_r

$\frac{n_i}{n}$ est la fréquence observée sur l'échantillon. Les écarts entre les fréquences observées $\frac{n_i}{n}$ et les probabilités p_i *peuvent-ils être attribués au hasard* ? Ou *sont-ils significatifs d'une différence réelle ce qui conduit au rejet de* H_0 *?*

Pour le vérifier, on pourrait utiliser la somme de la différence des écarts : $\sum_1^r (n_i - np_i)$. Or cette somme est nulle et ne peut pas servir. Pour résoudre cette difficulté, *on définit une variable aléatoire* :

$$D^2 = \sum_1^r \frac{(n_i - np_i)^2}{np_i} = \sum_1^r \frac{n_i^2}{np_i} - n\ ;$$

D^2 possède des *propriétés* particulièrement *simples*. Elle suit une loi de probabilité connue : la loi de χ^2 (présentée au *chapitre I*), caractérisée par *un nombre de degré de liberté (ddl)* ***égal*** *au nombre de catégories* ***moins*** *un* ***moins*** *le nombre de paramètres à ajuster*.

Remarques

☞ *Pourquoi moins 1 ?*

Il est évident que si l'on fixe 2 fréquences, la 3ième est immédiatement déterminée. On en conclut que si deux fréquences s'ajustent au modèle, il en est de même pour la troisième.

☞ A *chaque paramètre* ajusté, on *perd un degré* de liberté.

☞ Le χ^2 permet de tester la différence entre une distribution de fréquences observées et un modèle théorique. Ce test est appelé *test du Khi-deux*.

❸ ***Conditions d'application pour le χ^2***

- *L'échantillon* doit être de *grande taille* ($n \geq 30$)
- Les *fréquences théoriques* pour chaque classe, ne soient *pas trop petites* : On utilise une règle empirique : « *On applique la loi lorsque le nombre de fréquences en dessous de 5 n'excède pas 20 ou 25% du total et que les fréquences restent supérieures à 1.* » La fréquence théorique admise est donc 5, « *si le nombre de fréquences inférieures à 5 est trop élevé, on regroupe certaines catégories.* »

❹ ***Décision***

Si *l'hypothèse H_0* est vraie, la variable D^2 est distribuée comme une loi de χ^2 dont le *degré de liberté* est $(r - 1)$. On choisit alors un niveau de probabilité α auquel correspond une valeur de χ^2_α telle que :

$$Prob\ (\chi^2 \geq \chi^2_\alpha) = \alpha$$

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité α,* si pour un échantillon donné, la valeur trouvée dans l'expression D^2 *est inférieure à* χ^2_α

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***.

Remarques

☞ On peut choisir une infinité d'intervalles d'acceptation [a ; b] qui correspondent au même niveau α : Prob $(a < \chi^2 < b) = 1 - \alpha$. Dans ce cas pour *chaque intervalle*, il existe un test et il faut *choisir* celui pour lequel la *probabilité* d'accepter l'hypothèse, quand elle est fausse, soit *minimum*.

☞ Si l'hypothèse est vraie les n_i sont approximativement égaux aux np_i, sinon certains n_i auront une valeur différente des np_i et l'ordre de grandeur de l'expression D^2 est égal à n.

Dans la pratique :

◆ $np_i \geq 30$ la variable D^2 est une variable de χ^2 de *degré de liberté (r – 1)*.

◆ *Pour un degré de liberté* ≤ 30 les valeurs de χ^2 sont données par une table.

◆ *Pour un degré de liberté* > 30, on admet que $\sqrt{\chi^2}$ est une variable aléatoire approximativement *normale* de *moyenne* $\sqrt{2\nu - 1}$ et *d'écart-type 1*.

I – 1. 2 Applications

a) Test de proportions

❶ *Analyse*

Soit un événement A associé à une *expérience aléatoire* E. On répète n fois l'expérience. A apparaît k fois. On se propose de tester H_0 « l'événement a dans chaque épreuve une *probabilité d'apparition* $p = p_o$. » Il s'agit de décider si l'échantillon ne provient pas d'une population dont la distribution est une loi de *Bernoulli* de *paramètres 1* et $p = p_o$. On groupe les observations en deux classes suivant que l'événement est réalisé avec une probabilité $p = p_o$ ou non avec la probabilité $q = 1 - p$. Il vient :

$$\chi^2 = \frac{(k-np)^2}{np} + \frac{[(n-k)-nq]^2}{nq}$$

et en réduisant au même dénominateur, on obtient :

$\chi^2 = \dfrac{(k-np)^2}{npq}$ qui est le carré d'une variable centrée réduite

$T = \dfrac{(k-np)}{\sqrt{npq}}$. C'est une loi de *khi-deux* à *un degré* de liberté.

❷ *Décision*

On choisit alors un niveau de probabilité α, auquel correspond une valeur χ^2_α telle que : $Prob\ (\chi^2 \geq \chi^2_\alpha) = \alpha$

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si pour un échantillon donné, la valeur trouvée dans l'expression D^2 *est inférieure à* χ^2_α

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***.

❸ ***Correction de Yates***

Comme la loi de χ^2 est une distribution continue, il est préférable d'appliquer une correction de *continuité* ($\pm\frac{1}{2}$) dite *correction de Yates* et on calcule :

$$\chi_c^2 = \sum_1^r \frac{(\mid \nu_i - np_i \mid -0,5)^2}{np_i}$$

b) **Test de normalité**

Il s'agit de tester si la population ***P*** dans laquelle on a prélevé X est normale $N(m\ ;\ \sigma)$. ***Deux cas à étudier*** :

① *m et σ connus*

❶ ***Analyse***

En supposant les conditions d'application de χ^2 vérifiées, on considère r classes et on calcule $\chi^2_{\alpha,\ r-1}$

❷ ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si pour un échantillon donné, la valeur trouvée dans l'expression D^2 *est inférieure à* $\chi^2_{\alpha,\ r-1}$

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***.

② *m et σ inconnus*

❶ ***Analyse***

X suit une loi normale de *moyenne* et *d'écart-type inconnus*. A partir de l'échantillon, on *estime* ces *paramètres par la MMV* (méthode de maximum de vraisemblance). Si le nombre de paramètres à estimer $s < r$ avec r le nombre d'intervalles I_i, D^2 est une variable de khi-deux dont le degré de liberté est égal à $r - s - 1$ quand *l'hypothèse est vraie*.

❷ ***Décision***

Dans notre *cas* $s = 2$

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si pour un échantillon donné, la valeur trouvée dans l'expression D^2 *est inférieure à* $\chi^2_{\alpha,\ r-3}$

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***.

I - 2 TEST D'HOMOGENEITE

❑ On considère deux populations P_1 et P_2 partagées en deux classes sur la base du critère « *avoir le caractère A* »

❑ On prélève avec remplacement et de manière indépendante k *échantillons aléatoires* d'effectifs $n_1, \ldots, n_k$

❑ On observe $\nu_1, \ldots, \nu_k$ le nombre de fois de l'apparition de l'événement A

❑ On se propose de comparer $p_1 \ldots, p_k$

Analyse

Il s'agit de tester que les populations sont homogènes c'est à dire *qu'elles sont subdivisées dans les mêmes proportions.*

On teste *l'hypothèse nulle* : $H_0 : p_1 = \ldots = p_k = p$

par rapport à *l'hypothèse alternative* $H_1 : p_1 \neq \ldots \neq p_k \neq p$

Deux cas à distinguer selon la valeur de p

a) p connue

• Si les *effectifs sont petits*, on considère les k *suites* comme *une seule* avec :

$$\sum_1^k n_i = n \text{ et } \sum_1^k \nu_k = \nu \quad \text{et} \quad \chi^2 = \frac{(k-np)^2}{npq}$$

Lorsque n *est suffisamment grand* χ^2 *est* de *degré de liberté 1*

• Si les n_i *sont suffisamment grands*, on peut associer à chacune une variable de χ^2 de *degré 1* et $\chi_i^2 = \frac{(\nu_i - pn_i)^2}{pqn_i}$. Comme *khi-deux* est une *loi stable* alors $\chi^2 = \sum_1^k \chi_i^2$ est une variable de *khi-deux* de *degré de liberté k.*

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si

$\chi^2_{k-1} \leq \chi^2_{k-1,\ 1-\alpha}$

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***.

b) p inconnue

Il faut au préalable l'estimer par :

$$\hat{p} = \frac{v}{n} = \frac{v_1 + \ldots + v_k}{n_1 + \ldots n_k} \text{ et } \hat{q} = 1 - \hat{p}$$

Comme précédemment, on associe à chacune une variable de χ^2 de *degré 1*. Comme khi-deux est une loi stable alors $\chi^2 = \sum_1^k \chi_i^2$ est une *variable de khi-deux* de *degré de liberté* (*k-1*).

I – 3 TEST D' INDEPENDANCE

On se propose d'étudier deux variables X et Y dans une population ***P***. On tire un échantillon de taille n et on compare les *fréquences observées* avec les *fréquences théoriques espérées*. On regroupe, dans un tableau rectangulaire à double entrée, les observations en k *classes* pour X et r *classes* pour Y. On obtient le *tableau suivant* :

	Y_1		Y_j		Y_r	Σ
X_1	n_{11}		n_{1j}		n_{1r}	$n_{1.}$
X_i	n_{i1}		n_{ij}		n_{ir}	$n_{i.}$
X_k	n_{k1}		n_{kj}		n_{kr}	$n_{k.}$
Σ	$\boldsymbol{n_{.1}}$		$\boldsymbol{n_{.j}}$		$\boldsymbol{n_{.r}}$	$\boldsymbol{n..}$

❶ ***Analyse***

Hypothèse H_0 : « *Les deux variables X et Y étudiées simultanément dans la population* ***P*** *sont indépendantes* » : $p_{ij} = p_{i.}p_{.j} \quad \forall i \text{ et } j$

Comme les $p_{i.}$ et $p_{.j}$ ne sont pas connus, on les estime par leurs fréquences :

$$p_{i.} = \frac{n_{i.}}{n} \qquad \text{et} \qquad p_{.j} = \frac{n_j}{n}$$

d'où $v_{ij} = np_{ij}$

❷ ***Décision***

On utilise la statistique :

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(n_{ij} - \nu_{ij})^2}{\nu_{ij}}$$

✓ Si $n \geq 30$ et $\nu_{ij} \geq 5$, χ^2 suit une distribution de χ^2 *à* $(k-1)(r-1)$ *degrés de liberté*, si les variables sont indépendantes $\chi^2 = 0$

✓ Pour tester l'hypothèse H_0 contre H_1 , *au niveau* α, on utilise

$$\text{Prob}\,(\chi^2 > \chi^2_c / H_0) = \alpha$$

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si

$\chi^2_0 < \chi^2_c$; χ^2_c étant la *valeur lue dans la table* de distribution de *khi-deux à* $(k-1)(r-1)$ degrés de liberté et χ^2_0 celle *observée pour la statistique* définie dans *l'échantillon.*

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***

❸ ***Coefficient de contingence***

Lorsqu'il y a dépendance entre X et Y, on se demande souvent *quel est le degré* ou *l'ampleur de cette dépendance* ? Pour répondre à cette interrogation, on calcule le *coefficient de contingence de Cramer* défini comme suit :

$$C = \frac{\chi_0^2}{n(q-1)} \qquad 0 \leq C \leq 1$$

χ^2_0 la valeur *observée*, *q étant la valeur minimale entre k et r*

➢ Ce coefficient, *dépendant des dimensions k et r*, est à manier avec prudence !

II - TEST DES SIGNES

Si la distribution de la population *n'est pas normale* et si la *taille* des échantillons est *grande*, on utilise un *test non paramétrique* appelé ***test des signes***.

II - 1 TEST SUR UNE MOYENNE

Hypothèses

$$\begin{cases} H_0 & \mu = \mu_0 \\ H_1 & \mu < \mu_0 \qquad \mu > \mu_0 \qquad \text{ou} \qquad \mu \neq \mu_0 \end{cases}$$

❶ ***Analyse***

Si la distribution de la population est *symétrique*, la probabilité qu'une valeur d'un échantillon aléatoire soit inférieure à μ_0 (ou qu'elle soit supérieure ou égale à μ_0) est égale à $\frac{1}{2}$. Soit $x_1, \ldots, x_n$ les valeurs-échantillon de taille n.

◆ On remplace chaque valeur supérieure à μ_0 par (+) et chaque valeur inférieure par (-)

◆ On ne tient pas compte des valeurs égales

◆ On est ramené alors à un test sur une *proportion p*. En effet, on considère l'ensemble Ω des (+) et des (-) comme une variable de *Bernoulli* de paramètre p. Cette variable est définie comme étant la proportion des (+) dans l'ensemble Ω.

❷ ***Décision***

On teste alors :

$$\begin{cases} H_0 & p = \frac{1}{2} \\ H_1 & p < \frac{1}{2} \qquad p > \frac{1}{2} \text{ ou } \quad p \neq \frac{1}{2} \end{cases}$$

La comparaison du *nombre observé* avec la *valeur critique* permet le *choix* entre H_0 et H_1

II - 2 TEST DE COMPARAISONS

❶ ***Analyse***

Le test des signes est également utilisé lorsque les *différences* $D_i = X_i - Y_i$ des deux échantillons de taille n proviennent des deux caractéristiques X et Y qui ne sont *pas distribuées normalement*.

◆ On remplace alors chacune des paires (x_i, y_i) par :

(+) si $x_i > y_i$ et (-) si $x_i < y_i$ on ignore $x_i = y_i$

❷ ***Décision***

On teste alors :

$$\begin{cases} H_0 & p \geq \frac{1}{2} \\ H_1 & p < \frac{1}{2} \end{cases} \qquad \begin{cases} H_0 & p \leq \frac{1}{2} \\ H_1 & p > \frac{1}{2} \end{cases} \qquad \begin{cases} H_0 & p = \frac{1}{2} \\ H_1 & p \neq \frac{1}{2} \end{cases}$$

La comparaison du *nombre observé* avec la *valeur critique* permet le *choix* entre les deux hypothèses.

III - TEST DES RANGS

III - 1 ECHANTILLONS INDEPENDANTS

Test U de Mann-Whitney

❶ ***Analyse***

Il s'agit de comparer *deux proportions*:

◆ On tire un échantillon de taille n_1 dans la *première* population et un second de taille n_2 dans la *deuxième*

◆ On teste les *hypothèses* :

H_0 :« les *lots* proviennent de *populations identiques* »

H_1 :« les *lots* proviennent de *populations différentes* »

◆ On se base sur le rang des observations des deux échantillons ;

◆ On ordonne l'ensemble des observations en partant de la *plus petite* et on note son *rang* et sa *provenance* ;

◆ Lorsque les *observations sont égales*, on leur attribue *la moyenne des rangs* qu'elles auraient occupés si elles avaient été différentes.

❷ ***Décision***

H_0 vraie « on suppose que la somme des rangs du 1^{er} échantillon sera la même que la somme des rangs du $2^{ième}$ (les *deux provenant* de *populations identiques*) ».

◆ On utilise la statistique définie par :

$$Z = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \quad \text{avec } R_1 \text{ la somme des rangs du } 1^{er} \text{ échantillon.}$$

◆ On démontre que :

$$\overline{\mu}_z = \frac{n_1 n_2}{2} \quad \text{et} \quad \sigma_z = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$$

Si $n_1 \geq 8$ et $n_2 \geq 8 \Rightarrow$ Z suit *approximativement* une *distribution normale avec* $T = \dfrac{Z - \overline{\mu}_z}{\sigma_z} \rightarrow N(0, 1)$ *comme variable centrée réduite.*

Soit t_0 la valeur observée : $t_0 = \dfrac{z_0 - \overline{\mu}_z}{\sigma_z}$

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si

$t_0 \in [- t_{\alpha/2} ; t_{\alpha/2}]$

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***.

III - 2 ECHANTILLONS APPARIES

Test du signe de Wilcoxon

❶ ***Analyse***

On part des différences entre les mesures obtenues et on teste :

H_0 « *le signe de ces différences est du au hasard*»

❷ ***Décision***

Si H_0 est vraie les sommes (+) et (-) devraient être égales aux fluctuations d'échantillonnage près.

Pour de *faibles échantillons* ($n < 30$) le calcul s'obtient par la loi *binomiale*. Pour *n grand*, on utilise l'approximation de cette loi par la *loi normale*.

IV - TEST DES SEQUENCES

❶ ***Analyse***

Ce test des séquences (en anglais *runs test*) permet de vérifier si un échantillon est véritablement aléatoire *(i.i.d.)*.

◆ On appelle *séquence* une suite de lettres (ou d'autres symboles) identiques.

Exemple

On considère deux échantillons de tailles respectives n_1 et n_2 .
Si on désigne par X les *résultats* fournis par « le *premier échantillon* » et par Y les *résultats* fournis par « le *second échantillon* ». Si on considère la suite : X Y Y X X Y

séquence séquence ...

on *désigne par séquence,* une *suite de résultats* de *même famille*.

◆ On tire un échantillon de taille n, on note le nombre de symbole de la 1ère espèce, de la 2ème espèce et on définit la *statistique* :

U « *nombre total des séquences dans la série d'observations* » avec

$$\overline{U} = \frac{2n_1n_2}{n_1+n_2}+1 \qquad \text{et} \qquad \sigma_u = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

Si $n_1 \geq 10$ et $n_2 \geq 10$ $\Rightarrow U$ peut être considérée comme une *distribution normale de variable réduite* $T = \dfrac{U - \overline{U}}{\sigma_u} \rightarrow N(0, 1)$.

Il est préférable d'utiliser « la *correction de continuité* » soit

$$T = \frac{(U + \frac{1}{2}) - \overline{U}}{\sigma_u}$$

❷ ***Décision***

✧ ***L'hypothèse H_0*** est *acceptée, au niveau de probabilité* α, si

$t_0 \in [- t_{\alpha/2} ; t_{\alpha/2}]$

✧ ***L'hypothèse H_0*** est *rejetée* ***sinon***

CONCLUSION GENERALE

▪Comparés aux *tests paramétriques,* les *tests non-paramétriques* s'effectuent, également sur des *observations aléatoires* privilégiant *l'hypothèse nulle* H_0 .

▪Ils sont *moins contraignants* d'où leur utilité et leur développement.

▪Ils présentent néanmoins le *désavantage* d'être *moins puissants*.

▪Ils s'appliquent à la comparaison entre *deux échantillons* de *mesures ordinales*.

▪ D'autres *tests sont utilisés* pour la *comparaison entre plusieurs échantillons* :

- Le *test de Kruskall-Wallis* pour les *échantillons indépendants*
- Le *test de Friedman* pour les *échantillons appariés*.

La *procédure* de ces méthodes alternatives est *analogue* à celle que nous avons décrite pour le *test des rangs*.

Auto-évaluation ?

1- On considère k suites d'observations indépendantes, d'effectifs respectifs n_1 ;... ; n_k. On désigne par $\gamma_1, \ldots, \gamma_k$ la réalisation d'un événement A. On se propose de tester l'hypothèse H_0 « *l'évènement a dans chaque suite une probabilité constante p* ». Pour appliquer le test de χ^2, on peut

a) réunir toutes les suites en une seule

b) associer à chacune des séries une variable χ^2

c) associer à chacune des séries une variable χ^2 , lorsque tous les n_i sont suffisamment grands.

2- Pour rejeter H_0 il faut que :

a) l'une des valeurs calculées soit supérieure à la valeur lue

b) toutes les valeurs calculées soient supérieures à la valeur lue.

3- Pour tester l'hypothèse H_0 « *X est une variable aléatoire ayant pour fonction de répartition F(x)* », on peut subdiviser le domaine de variation de X en intervalles finis $I_1, \ldots, I_r$

a) de longueurs constantes b) de longueurs quelconques.

4- Les probabilités qui correspondent aux intervalles I_r sont données par :

a) les observations

b) la fonction de répartition $F(x)$.

5- Si l'hypothèse H_0 est vraie, χ^2 est de degré :

a) $n - 1$ b) $r - 1$

6- Est-ce que le test d'ajustement porte sur :

a) $F(x)$

b) toute fonction de répartition, donnant des probabilités $p_1, \ldots, p_r$

7- Pour un test d'ajustement , on choisit un test

a) bilatéral b) unilatéral

8- En pratique, lorsqu'on admet que la variable D^2 est de degré (r –1), les

a) $np_i \geq 5$ b) np_i quelconques

9- Les valeurs de χ^2 sont fournies par les tables :

a) quel que soit le degré de liberté ν

b) Pour les valeurs de $\nu \leq 30$ seulement.

10- Lorsque la fonction de répartition dépend de plusieurs paramètres inconnus, il faut :

a) les estimer d'abord

b) appliquer directement le test du χ^2

11- Dans ce dernier cas, le nombre de degré de liberté est :

a) $r - 1$ b) $r - s - 1$

12- Pour une même hypothèse, on peut considérer :

a) un seul test correspondant au niveau α

b) plusieurs tests correspondant au niveau α

13- Dans un test de comparaisons d'échantillons appariés, les séries de mesure sont :

a) indépendantes

b) ne sont pas indépendantes.

14- Si le nombre de couples d'observations $n > 30$, on peut utiliser la loi normale centrée réduite lorsque la population des différences est distribuée selon :

a) une loi normale

b) quelle que soit la loi

15- Le test du signe de Wilcoxon est utilisé pour les échantillons

a) indépendants

b) appariés.

16- Le test U de Mann-Whitney est utilisé pour les échantillons :

a) indépendants

b) appariés.

Exercices ?

Exercice 1

Supposons que dans une population, il y ait une équipartition entre les hommes et les femmes. On tire un échantillon de 200 personnes et on trouve 120 de sexe féminin. Comparer la distribution selon le sexe.

Exercice 2

Une enquête par sondage sur la population active dans la région de Tunis a donné les résultats suivants :

	Personnes interrogées	Cadres moyens
Hommes	3000	390
Femmes	2000	228

Le facteur sexe a-t-il une influence d'être un cadre, au seuil $\alpha = 0,15$?

Exercice 3

Une enquête par sondage sur les remises consenties sur les achats des ordinateurs à Tunis et ailleurs a fait apparaître les résultats suivants :

Achats / Lieu	Avec remise	Sans remise	Total
Tunis	26	34	60
Ailleurs	26	34	60
Total	52	68	120

En prenant un risque de 1ère espèce $\alpha = 0,15$, que peut-on conclure ?

Exercice 4

Soit X le nombre d'accidents qui se produisent par mois, à un carrefour déterminé de Tunis. L'observation de X pendant 2 ans a fourni la répartition suivante :

X	0	1	2	3	4	5	6	7
n_i	1	3	8	10	5	2	0	1

1) Calculer la moyenne et l'écart-type de cette distribution.
2) L'ajustement à une loi de Poisson de paramètre $m = 3$ de cette série statistique est-il acceptable avec un risque de première espèce de10% ?

Exercice 5

La répartition du personnel d'une entreprise selon la classe salariale est donnée par le tableau suivant :

X	[0-10[	[10-20[	[20-30[	[30-40[	[40-50[	[50-60[	[60-70[	[70-80[	[80-90[
n_i	5	6	40	168	288	277	165	49	2

(le salaire est exprimé en dizaine de dinars)

1) Calculer la moyenne et l'écart-type de cette distribution.
2) L'ajustement à une loi normale d'espérance 50 et d'écart-type 13 de cette série statistique est-il acceptable avec un risque de première espèce de 5 % ?

Exercice 6

Une enquête parmi les conducteurs d'un certain type de voiture a montré la répartition suivante du nombre d'accidents par an, en fonction de l'âge du conducteur :

X / Y	]18-25]	]25-50]	> 50	*Total*
[0-2[	23	54	16	93
≥ 3	22	21	14	57
Total	45	75	30	150

Peut-on avec un risque de 1ère espèce de 5 % en déduire que l'âge du conducteur a une influence sur le nombre annuel d'accidents ?

Exercice 7

Un marchand de journaux désire optimiser sa stratégie. Il achète ses journaux à 0,3 D pièce et les vend à 0,4 D. On lui reprend les invendus à 0,05 pièce. Pendant 100 jours où il a eu sans cesse des invendus, il a noté le

nombre de journaux vendus. La moyenne de l'échantillon et l'écart-type sont calculés d'après les données suivantes :

$$\sum_{1}^{100} x_i = 25000 \quad \text{et} \quad \sum_{1}^{100} (x_i - 250)^2$$

1) Estimer la moyenne et l'écart- type de la population.
2) En considérant la distribution suivante :

X	$-\infty$;175	175-200	200-225	225-250	250-275	275-300	300-325	325 ; $+\infty$
n_i	8	8	12	22	17	18	10	5

Tester à 5 % par le test du χ^2 l'ajustement à la loi normale *N(250 ; 50)*

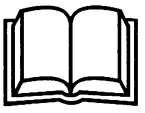
3) Donner un intervalle de confiance à 95 % de m.
4) Quel doit être le nombre S de journaux que le marchand doit avoir, au début de la journée, pour maximiser son profit ?

Exercice 8

La distribution du nombre de consultants de manuels de statistiques, à la bibliothèque de la Faculté de médecine de Tunis, durant une semaine ordinaire de l'année, est donnée par le tableau suivant :

Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Σ
8	15	22	20	13	12	90

Peut-on, avec un risque de 5%, en déduire que les consultations de manuels de statistiques ne dépendent pas du jour de la semaine ?



Corrigés

Auto-évaluation

1- a; c
2- a ; b
3- b
4- b
5- b
6- b
7- b
8- b
9- b
10- a
11- b
12- b
13- b
14- a; b
15- b
16- a

Exercices

Exercice 1

	F	M	Total
Effectifs théoriques	100	100	200
Effectifs observés	120	80	200
Total	220	180	400

Calcul des effectifs théoriques

	F	M	Total
Effectifs théoriques	110	90	200
Effectifs observés	110	90	200
Total	220	180	400

$$\delta^2 = 10^2\left(\frac{1}{110}+\frac{1}{90}+\frac{1}{110}+\frac{1}{90}\right) = 4{,}04$$

Règle de décision :

$\begin{cases} \delta^2 \geq 3{,}841 \Rightarrow & \text{dépendance de l'ajustement} \\ \delta^2 < 3{,}841 \Rightarrow & \text{non significatif risque de } 1^{\text{ère}} \text{ espèce } \alpha = 0{,}05 \end{cases}$

Comme la valeur calculée est supérieure à 3,841, l'enquête a mis en évidence l'influence du sexe.

Exercice 2

En désignant par CM les cadres moyens et par PA la population active interrogée, on obtient le tableau suivant :

	CM	PA	Total
H	390	2610	3000
F	228	1772	2000
Total	618	4382	5000

On considère l'hypothèse H_0 :

« *Les deux variables X* (hommes) et *Y* (femmes) sont *indépendantes* »

$$n'_{ij} = \frac{n_i \, n_j}{n}$$

On calcule les effectifs théoriques comme suit :

$n'_{11} = \frac{3000 \times 618}{5000} = 370,8$ ....les autres effectifs se calculent par différences :

	CM	Autres	Total
X	370,8	2629,2	3000
Y	247,2	1752,8	2000
Total	618	4382	5000

Tous les effectifs sont supérieurs à 5 donc

$$D^2 = \sum \frac{(n'_{ij} - n_{ij})^2}{n'_{ij}} \rightarrow \chi^2 \text{ de degré de liberté 1.}$$

La *table du khi-deux* donne pour $\alpha = 0,15$ la valeur *2,17*

Règle de décision :

$\begin{cases} \delta^2 \geq 2,17 \Rightarrow \text{dépendance de l'ajustement} \\ \delta^2 < 2,17 \Rightarrow \text{on ne peut affirmer une influence du facteur sexe sur la situation.} \end{cases}$

Exercice 3

Calcul des effectifs théoriques :

	Avec remise	*Sans remise*	Total
Tunis (*X*)	26	34	60
Ailleurs (*Y*)	26	34	60
Total	52	68	120

$$\delta^2 = 4^2\left(\frac{1}{26} + \frac{1}{26} + \frac{1}{34} + \frac{1}{934}\right) = 2,17$$

On effectue les *mêmes calculs* que dans l'exercice *(2)*

Règle de décision :

$\begin{cases} \delta^2 \geq 3,841 \Rightarrow \text{dépendance de l'ajustement} \\ \delta^2 < 3,841 \Rightarrow \text{non significatif risque de } 1^{\text{ère}} \text{ espèce } \alpha = 0,05 \end{cases}$

Comme la valeur calculée est inférieure à 3,841, l'enquête n'a pas mis en évidence l'influence du lieu géographique sur les fréquences d'achat avec remise.

Exercice 4

① *Calcul de la moyenne et de l'écart- type de la distribution* :

X	0	1	2	3	4	5	6	7	Total
n_i	1	3	8	10	5	2	0	1	30
$n_i x_i$	0	3	16	30	20	10	0	7	86
$n_i x^2_i$	0	3	32	90	80	50	0	49	304

On calcule $m = \frac{86}{30} = 2{,}867$ et $\sigma^2 = \frac{304}{30} - (2{,}867)^2 \# 2{,}49$

② Ajustement à une loi de *Poisson* de paramètre $m = 3$ avec un risque de *première espèce* de *10 %.*

X	0	1	2	3	4	5	6	7	Total
n_i	1	3	8	10	5	2	0	1	30
p_i	0,0498	0,1494	0,224	0,224	0,168	0,1008	0,0504	0,0336	1

Le calcul des np_i conduit à regrouper les deux premières et les trois dernières classes en une classe *supérieure ou égale à 5* d'où le tableau suivant :

np_i	5,98	6,72	6,72	5,04	5,54	Total
δ^2	0,656	0,244	2,726	0	2,262	5,89

Règle de décision : Par lecture dans la table du $\chi^2_4 = 7{,}78$.

$\begin{cases} \delta^2 \geq 7{,}78 \Rightarrow \text{on refuse l'ajustement} \\ \delta^2 < 7{,}78 \Rightarrow \text{on accepte l'ajustement.} \end{cases}$

Le calcul *donne* $\delta^2 = 5{,}89 \Rightarrow$ on *accepte l'ajustement* de la distribution à une loi de *Poisson de paramètre* $m = 3$.

Exercice 5

① *Calcul de la moyenne et de l'écart- type de la distribution* :

$\bar{x} = 49{,}76$ et $S = 12{,}8$

② *Ajustement à une loi normale*
En utilisant la variable centrée réduite :

$$t_i = \frac{x_i - 50}{13}$$

X	n_i	t_i	F	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
0- 10	11	-3,85	-	0,00097	10,44	0,03
10-20		-3,08	0,00097	0,00947		
20-30	40	-2,31	0,01044	0,05134	51,34	2,5
30-40	168	-1,54	0,06178	0,15887	158,87	0,525
40-50	288	-0,77	0,22065	0,27935	279,35	0,268
50-60	277	0	0,5	0,27935	279,35	0,02
60-70	165	0,77	0,77935	0,15887	158,87	0,236
70-80	49	1,54	0,93822	0,05134	51,34	0,107
80-90	2	2,31	0,98956	0,01044	10,44	6,823
Total	1000			1	1000	# 10,51

On dispose de *sept degrés de liberté*, on lit la valeur correspondante dans la table, on obtient $\chi^2_7 = 14,07$
Règle de décision :

$\delta^2 \geq 14,07 \Rightarrow$ on refuse l'ajustement
$\delta^2 < 14,07 \Rightarrow$ on accepte l'ajustement

Le calcul précédent nous *donne* $\delta^2 = 10,51 \Rightarrow$ on *accepte l'ajustement* de la distribution à une loi normale *N(50 ; 13)*.

Exercice 6

On considère l'hypothèse H_0 « les deux variables X (âge) et Y (nombre d'accidents) sont *indépendantes* ».
On calcule les effectifs théoriques :

X / Y	]18-25]	]25-50]	> 50	Total
[0 ; 1 ; 2]	27,9	46,5	18,6	93
≥ 3	17,1	28,5	11,4	57
	45	75	30	150

Le calcul donne :

$$\delta^2 = \frac{(4,9)^2}{27,9} + \frac{(4,9)^2}{17,1} + \frac{(7,9)^2}{46,5} + \frac{(7,5)^2}{28,5} + \frac{(2,6)^2}{18,6} + \frac{(2,6)^2}{11,4} = 6,405$$

On dispose de *deux degrés de liberté*, on lit la valeur correspondante dans la table, on obtient $\chi^2_2 = 5,991$

Règle de décision :

$\begin{cases} \delta^2 \geq 5,991 \Rightarrow & \text{on rejette } H_0 \text{ on admet l'influence de l'âge} \\ \delta^2 < 5,991 \Rightarrow & \text{on ne peut rejeter } H_0 \text{ les variations ne sont pas significatives} \end{cases}$

Le calcul précédent nous *donne* $\delta^2 = 6,405 > 5,991 \Rightarrow$ on peut conclure que le *facteur âge influe* sur le nombre d'accidents.

Exercice 7

① Les estimateurs sont : $\bar{x} = \frac{25000}{100} = 250$ et S = 50

L'estimateur sans biais est $S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

② La table *N(0 ; 1)* donne G(0) = 0,5 ;

G(0,5) = 0,6915 t = 0,1915

G(1) = 0,8413 t = 0,1498

G(1,5) = 0,9332 t = 0,0919

G(∞) = 1 t = 0,0668

X	-∞ -175	175-200	200-225	225-250	250-275	275-300	300-325	325- +∞
n_i	8	8	12	22	17	18	10	5
$p_i\ n_i$	6,7	9,2	15	19,1	19,1	15	9,2	6,7
δ^2	0,25	0,15	0,6	0,44	0,23	0,6	0,07	0,43

On dispose de *cinq degrés de liberté*, on lit la valeur correspondante dans la table, on obtient $\chi^2_5 = 11,1$

Règle de décision :

$\begin{cases} \delta^2 \leq 11,1 \Rightarrow & \text{l'hypothèse } H_0 \text{ est acceptée} \\ \delta^2 > 11,1 \Rightarrow & \text{l'hypothèse } H_0 \text{ est rejetée} \end{cases}$

Le calcul précédent nous *donne* $\delta^2 = 2,77 < 11,1 \Rightarrow$ l'hypothèse peut donc être acceptée.

③ *Intervalle de confiance pour m*

$\bar{x} \to N(m, \frac{\sigma}{\sqrt{n}})$ et $\bar{x} - \frac{2\sigma}{10} \leq m \leq \bar{x} + \frac{2\sigma}{10}$ d'où $240 < m < 260$

④ Soit S le nombre optimal. Une augmentation d'une unité va se traduire soit par

* $x \leq S$ ce qui correspond à une dépense de 0,250.

La probabilité dans ce cas est la probabilité pour $x \leq S$ soit G(S). L'espérance de perte supplémentaire est alors : 0,25 G(S)

* ou bien $x > S$ auquel cas une augmentation de 0,10 et la probabilité de $1 - G(S)$. L'espérance de gain sera alors : [0,1 (1 - G(S)]. Le stock optimal est tel que : 0,25 G(S) = [0,1 (1 - G(S)] d'où $G(S) = 0,285$. A cette valeur correspond dans la table $N(0 ; 1)$ un nombre $t = - 0,57$ et le stock optimum est donné par $S = 250 - 0,57 \times 50 = 222$.

Exercice 8

On considère l'hypothèse H_0 « chaque jour la probabilité de consultation est égale à $\frac{1}{6}$ ». Le test à effectuer est un test de khi-deux d'ajustement. L'effectif total étant égal à 90, les effectifs théoriques sont égaux à :

$$\frac{1}{6} \times 90 = 15 \qquad \text{d'où}$$

$$\chi^2_{obs} = \frac{(8-15)^2}{15} + \ldots + \frac{(12-15)^2}{15} = 8,4$$

On dispose de *cinq degrés de liberté*, on lit la valeur correspondante dans la table, on obtient $\chi^2_5 \# 11,1$ *(pour* $\alpha = 0,05$*)*

Règle de décision :

$\begin{cases} \delta^2 \leq 11,1 \Rightarrow & \text{l'hypothèse } H_0 \text{ est acceptée} \\ \delta^2 > 11,1 \Rightarrow & \text{l'hypothèse } H_0 \text{ est rejetée} \end{cases}$

Le calcul précédent nous *donne* $\delta^2 = 8,4 < 11,1 \Rightarrow$ l'hypothèse peut donc être acceptée, au risque de 5 %.

ANNEXES

SYNTHESE DES LOIS USUELLES

BIBLIOGRAPHIE SOMMAIRE

DES NOMS & DES DATES

INDEX ALPHABETIQUE

TABLES STATISTIQUES

SYNTHESE
DES LOIS USUELLES

I - Loi de BERNOULLI : $B(1, p)$

Une variable aléatoire X suit une loi $B(1, p)$ si elle prend les deux seules valeurs 0 et 1 avec les probabilités suivantes :

$$p\{X = 1\} = p$$
$$p\{X = 0\} = 1 - p$$

La loi s'écrit : $P\{X = x\} = p^x (1-p)^{1-x}$

❶ ***Espérance et Variance :***

$$E(X) = p \qquad V(X) = p(1-p)$$

On utilise une telle v.a lorsque les *résultats possibles* d'une épreuve aléatoire sont *réduits à deux* :

(*oui – non* ; *vrai – faux* ; *possède une propriété – ou non* ...)

II - Loi BINOMIALE : $B(n, p)$

Une v.a X suit une loi binomiale $B(n, p)$ si elle prend les $(n+1)$ valeurs $0,1,2,...,n$ avec les probabilités :

$$P\{X = x\} = C_n^x p^x (1-p)^{n-x} \qquad (x = 1,2, ...,n)$$

❶ ***Espérance et Variance***

$$E(X) = np \qquad V(X) = np(1-p)$$

❷ ***Construction***

Si n *v.a indépendantes* $X_1, X_2, ..., X_n$ suivent la même loi de Bernoulli $B(1, p)$ alors la v.a : $X = \sum_{i=1} X_i$ suit une loi Binomiale $B(n, p)$

❸ ***Addition***

Si 2 *variables aléatoires indépendantes* X_1 *et* X_2 suivent respectivement $B(n_1, p)$ et $B(n_2, p)$, alors $(X_1 + X_2)$ suit une *loi Binomiale* $B(n_1 + n_2, p)$

❹ ***<u>Approximation</u> (convergence en loi)***

Lorsque n est grand :

① si $n > 30$; si np *et* $np(1-p)$ sont voisins et si $np \leq 15$, on peut alors *approximer* la loi $B(n, p)$ par une *loi de Poisson* $P(\lambda)$ où $\lambda = np$

② si $n\,p > 15$ et $n\,(1-p) > 15$ (ou si $n\,p\,(1-p) > 5$) on peut *approximer* la loi B (n, p) par une *loi normale* $N\,(m, \sigma)$ où $m = n\,p$ et $\sigma = \sqrt{n\,p\,(1-p)}$.

Deux types d'approximation

➢ *l'une par une loi discrète* (ce sera souvent le cas pour des événements « *rares* » : *p très petit*),

➢ *l'autre par une loi continue*.

Les conditions de validité des approximations dépendent de la précision exigée.

III - Loi HYPERGEOMETRIQUE : *H(N, n, p)*

Une variable aléatoire X suit une loi $H(N, n, p)$ *si elle prend les valeurs* $0, 1, 2 \ldots n$ avec $Max\,(0, n - N_2) \leq x \leq Min\,(n, N_1)$

N : nombre d'individus de la population P

N_1 : nombre d'individus de la population P_1 et $p = \frac{N_1}{N}$

N_2 : nombre d'individus de la population P_2 *et* $q = \frac{N_2}{N}$ *et* $p + q = 1$

n : nombre d'individus choisis au hasard dans la population P avec $n \leq N$

tirage sans remise avec les probabilités

$$P\,(X = x) = \frac{C_{Np}^{x} C_{Nq}^{n-x}}{C_N^n}$$

❶ ***Espérance et Variance***

$E(X) = n\,p$ $\qquad V(X) = \frac{N-n}{N-1} npq$

$\frac{N-n}{N-1}$ *est appelé facteur d'exhaustivité*

❷ ***Approximation***

Lorsque $N > 10\,n$ (la taille de la population est 10 fois plus grande que celle de l'échantillon), on peut *approximer la loi hypergéométrique* $H\,(N, n, p)$ par une *loi Binomiale* $B\,(n, p)$.

IV - Loi de POISSON : $P(\lambda)$

Une variable aléatoire X suit une loi de Poisson $P(\lambda)$ si elle prend ses valeurs dans N avec les probabilités :

$$P\{X = x\} = e^{-\lambda} \frac{\lambda^x}{x!} \qquad x \in N \qquad \lambda > 0$$

❶ ***Espérance et variance***

$$E(X) = \lambda \qquad V(X) = \lambda$$

❷ ***Addition***

Si *2 variables aléatoires indépendantes* suivent respectivement les lois $P(\lambda_1)$ et $P(\lambda_2)$, alors la v.a.$(X_1 + X_2)$ suit une *loi de Poisson* $P(\lambda_1 + \lambda_2)$

V - Loi UNIFORME : $U(a, b)$

Une variable aléatoire suit la loi uniforme sur l'intervalle[a, b] si X est une variable aléatoire continue de *densité f*

$$f(x) = \frac{1}{b-a} \qquad si \qquad x \in [a, b]$$

$$f(x) = 0 \qquad si \qquad x \notin [a, b]$$

(la densité est constante sur l'intervalle $[a, b]$)

❶ ***Espérance et variance***

$$E(X) = \frac{a+b}{2} \qquad V(X) = \frac{(b-a)^2}{12}$$

VI - Loi de LAPLACE-GAUSS ou Loi NORMALE : $N(m, \sigma)$

Une variable aléatoire X suit une loi $N(m, \sigma)$ si elle est *continue* et de *densité f :*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} \qquad x \in R$$

❶ ***Espérance et variance***

$$E(X) = m \qquad V(X) = \sigma^2$$

❸ ***Lecture de table***

Soit une variable aléatoire X de loi $N(m, \sigma)$, X peut se mettre sous la forme suivante : $X = \sigma T + m$ où T suit une loi $N(0, 1)$. Cette dernière loi étant tabulée.

On a donc $P\{X < x\} \Leftrightarrow P\left\{T < \frac{x-m}{\sigma}\right\}$

❹ ***Addition***

Si *2 variables aléatoires indépendantes* suivent respectivement des lois normales $N(m_1, \sigma_1)$ et $N(m_2, \sigma_2)$ la *variable aléatoire* $(\mathbf{X_1} + \mathbf{X_2})$ suit une loi $N(m_1 + m_2, \sqrt{\sigma_1^2 + \sigma_2^2})$

❺ ***Combinaison linéaire***

On démontre que si $X_1, X_2, \ldots, X_n$ suivent des lois normales, pour tous $\lambda_1, \lambda_2, \ldots, \lambda_n \in \boldsymbol{R}$ $\sum_{i=1}^{n} \lambda_i X_i$ suit une loi normale.

VII - Lois DERIVEES de La LOI NORMALE

Certaines lois peuvent être construites à partir de la loi de Gauss

1 - Loi de Khi-deux : χ^2_n (*Khi-deux à n degrés de liberté*)

soit $U_1, U_2, \ldots, U_n$ *n v.a. indépendantes de même loi N(0, 1)* la v.a.

$$Y = \sum_{i=1}^{n} U_i^2 \quad \text{suit une loi de } \chi^2_n$$

❶ ***Espérance et variance***

$$E(\chi^2_n) = n \qquad V(\chi^2_n) = 2n$$

❷ ***Addition***

Si Y_1 et Y_2 sont 2 v.a. indépendantes et de loi respectives $\chi^2_{n_1}$ et $\chi^2_{n_2}$, la v.a. $Y_1 + Y_2$ suit une loi $\chi^2_{n_1+n_2}$.

❸ ***Approximation***

Lorsque n est *grand*, $n > 30$, la loi de la v.a. $\sqrt{2\chi^2_n} - \sqrt{2n-1}$ peut être approximée par une loi normale $N(0, 1)$.

2 - Loi de STUDENT : t_n (***Student à n degrés de liberté***)

Soit X et Y *2 v.a. indépendantes* telles que X suit une loi $N(0, 1)$ et Y suit une loi χ^2_n alors la v.a. $\frac{X}{\sqrt{\frac{Y}{n}}}$ suit une loi t_n

$$E(t_n) = 0$$

❶ ***Approximation***

Lorsque n est *grand*, n > 30, la loi t_n peut être approximée par une loi *normale* $N(0, 1)$.

3 – Loi de FISHER- SNEDECOR : $F(n_1, n_2)$

Soit X_1 *et* X_2 *deux v.a. indépendantes* de lois respectives $\chi^2_{n_1}$ et $\chi^2_{n_2}$.

Alors la variable aléatoire $\frac{\frac{X_1}{n_1}}{\frac{X_2}{n_2}}$ suit une loi $F(n_1, n_2)$ à n_1 *et* n_2 degrés de liberté.

VIII – STATISTIQUES DES ECHANTILLONS

1 - *Notation*

Si la variable aléatoire X suit la loi L, on notera $X \to L$

2 - *Définition d'un échantillon*

$X_1, X_2, \ldots, X_n$ est un échantillon n de X *si les variables aléatoires* $X_1, \ldots, X_n$ sont *indépendantes* et suivent la *même loi* que X.

3 - *Définition d'une statistique*

Une statistique est une fonction réelle des variables aléatoires de l'échantillon, c'est une variable aléatoire .

Soit alors $X_1, X_2, \ldots, X_n$ n-échantillon de X où $X \to N(m, \sigma)$

On a les *résultats suivants* :

- $\frac{\sum_{i=1}^{n} X_i}{n} \rightarrow N(m, \frac{\sigma}{\sqrt{n}})$

- $\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sigma^2} \rightarrow \chi^2_{n-1}$

car $(X_i - \overline{X})$ *ne sont pas indépendantes*

- $\frac{\overline{X} - m}{\frac{s'}{\sqrt{n}}} \rightarrow t_n$ où $s'^2 = \frac{\sum_{i=1}^{n} (X_i - m)^2}{n}$

- $\frac{\overline{X} - m}{\frac{S}{\sqrt{n}}} \rightarrow t_{n-1}$ où $S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$

Soit $(X_{n1}, ..., X_{n2})$ et $(Y_{n1}, ..., Y_{n2})$ deux *échantillons indépendants* de *lois respectives* $N(m_1, \sigma_1)$ et $N(m_2, \sigma_2)$:
On a les résultats suivants :

- $\frac{\sum_{i=1}^{n1} (X_i - m_1)^2}{n_1 \sigma_1^2} \Big/ \frac{\sum_{j=1}^{n2} (Y_j - m_2)^2}{n_2 \sigma_2^2} \rightarrow F(n_1, n_2)$

- $\frac{\sum_{i=1}^{n1} (X_i - \overline{X})^2}{(n_1 - 1)\sigma_1^2} \Big/ \frac{\sum_{j=1}^{n2} (Y_j - \overline{Y})^2}{(n_2 - 1)\sigma_2^2} \rightarrow F(n_1 - 1, n_2 - 1)$

- $\frac{(\overline{Y} - \overline{X}) - (m_2 - m_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow N(0, 1)$

BIBLIOGRAPHIE

Calot G. *Cours de statistique descriptive*, Dunod, 1981.
Cours de calcul des probabilités, Dunod, 1967.

Desabie *Théorie et pratique des sondages*, Dunod, 1971

Fourgeaud C. **F**uchs A. *Statistique*, Dunod, 1967.

Giard V. *Statistique appliquée à la gestion*, Economica, 1992.

Goldfard B. Pardoux C . *Introduction à la méthode statistique*, Dunod, 1998.
Grais B. *Méthodes statistiques*, Coll. Economie Module, Dunod, 1992.

Labrousse C. *Statistique*, Dunod, 1977.

Lebart L. **M**orineau A. **F**énelon J.P. *Traitement des données statistiques*, Dunod, 1979

Lecoutre J.P. *Statistique et probabilités*, Dunod, 1998.

Levy M. *L'informatique statistique*, Seuil 1975.

Martel J.M. **N**adeau R. *Statistique en gestion et en économie*, gaetan Morin éditeur, Canada, 1988.

Mothès J. Prévisions et décisions statistiques dans l'entreprise, Dunod, 1968
Pasquier A. *Statistique*, Dalloz, 1971.

Saporta G. *Probabilités, analyse des données et statistique*, Technip, 1990

Wonnacott T, Wonnacott R. *Statistique*, Economica, 1991.

DES NOMS & DES DATES

*A*l Khawarizmi Mohamed Ibnou Moussa (780-874) mathématicien et astronome musulman considéré comme l'un des meilleurs mathématicien, a fondé l'algèbre et le mot Algorithme.

*B*ernoulli famille de savants: mathématiciens, astronomes et physiciens suisses. Les plus célèbres:

*J*acques I (1654-1705) a développé le calcul différentiel et intégral en Europe;

*D*aniel (1700-1782) fils de Jean a travaillé sur les probabilités et a été le fondateur de la physique mathématique.

*B*oole George(1815-1864) mathématicien anglais, fondateur de l'algèbre de la logique.

*C*ournot Antoine Augustin (1801-1877) mathématicien, économiste et philosophe français, connu pour ses travaux sur le calcul des probabilités et pour sa théorie du hasard.

*F*ermat Pierre (1601-1665) mathématicien français a établi les bases du calcul infinitésimal et du calcul des probabilités.

*G*alton Francis (1822-1911) naturaliste anglais cousin de Darwin a appliqué les méthodes statistiques à la physiologie et à l'hérédité.

*G*auss Carl Friedrich (1777-1855) mathématicien, physicien et astronome allemand. Il a inventé la méthode des moindres carrés.

*G*uieth el Keshi (1359-1436) inventeur des fractions décimales et du périmètre. Son livre « miftah el hissab » (la clé des mathématiques) est considéré comme une encyclopédie des mathématiques.

Kolmogorov Andrei Nikolaievitch (1903-1987) mathématicien russe, connu pour sa théorie axiomatique des probabilités.

Laplace Pierre Simon (marquis)(1749-1827) mathématicien physicien et astronome français, connu par sa théorie analytique des probabilités

Laspeyres et ***P***aasche statisticiens allemands, ont proposé respectivement en 1864 et 1874 les formules des indices synthétiques.

Markov Andrei (1856-1936) mathématicien russe a introduit les chaînes d'événements et a modernisé le calcul des probabilités.

Moivre Abraham(de) (1667-1754) mathématicien anglais spécialiste du calcul des probabilités

Pascal Blaise (1623-1662) savant, philosophe et écrivain français. Il a jeté les bases de calcul des probabilités et a étudié l'analyse combinatoire.

Pearson Karl (1857-1936) mathématicien britannique a inventé le test de khi-deux.

Poisson Siméon Denis (1781-1840) mathématicien et homme politique français. Connu pour ses travaux sur le calcul des probabilités, sur ses travaux de mécanique rationnelle et sur l'application des probabilités dans les domaines du droit.

Pythagore (VI^{e} siècle av.J.C.) philosophe et mathématicien grec, créateur des sciences mathématiques.

Quételet Adolphe (1796-1874) astronome et mathématicien belge. Ses recherches appliquées à la taille des hommes sont à l'origine de l'anthropométrie.

Student pseudonyme de William Sealy ***G***osset (1876-1937) industriel (brasserie) et scientifique.

INDEX ALPHABETIQUE

A

*A*baque, 104
*A*cceptation,
    région d' 136, 140
*A*justement
    test, 211
*A*léatoire, 7, 45, 52, 55, 161
    échantillon, 46, 49, 54
*A*lternative, 140
*A*nalyse, 149
    à priori, 138
    à posteriori, 138
    bayesienne, 137
    statistique, 38
*A*NOVA, 159
*A*pproximation, 18, 142
*A*platissement
    coefficient, 6
*A*ppariés, 107, 113, 148, 159

B

Base, 45
Bayes, 138
Bernoulli, 12, 81, 88
Biais, 43, 73, 153
Bilatéral, 142, 147
Binomiale, 77, 146
Bienaymé-Tchebychev, 9, 72

C

*C*oefficient, 97
    d'aplatissement, 6
    d'asymétrie, 6
    de corrélation, 118,164

*C*onfiance 43, 102
intervalle de 102
*C*onvergence, 6
en loi, 7
en probabilité, 8
en moyenne quadratique, 8
*C*ontingence
coefficient, 197
tableau, 148, 196
*C*onvergent, 71
*C*orrection, 194, 201
*C*orrélation, 102, 118
*C*ramer, 75, 197
*C*ritique
région, 136, 139
valeur, 164, 198

D

Décision, 39, 138, 144
Degrés de liberté, 18, 19, 48, 161, 191
Dénombrement, 39
Densité de probabilité, 10, 17
Dérivées, 12, 80
Différences, 199
Dispersion, 5
Distribuées (i.i.d.), 70
Distribution, 81

E

Ecart absolu moyen, 4
Ecart-type, 5, 54
Echantillon
aléatoire, 38, 53
apparié, 109
grand, 158
taille d'un, 41, 46, 108
simple avec ou sans remise, 52, 45

*E*chantillonnage, 39, 53
*E*fficace, 84,
estimateur, 75, 81, 164
*E*fficacité, 75
*E*llipse, 104
*E*mpirique, 42
*E*quiprobable, 3
*E*rreurs de type I et II 137
*E*stimation, 71
*E*stimateur, 70, 167
convergent, 71, 167
des moindres carrés, 79
du maximum de vraisemblance, 84
efficace, 74
exhaustif, 76
*E*xhaustif, 45
*E*xponentielle, 15, 88

F

*F*acteur
un seul, 159
deux facteurs, 161
*F*acteur d'exhaustivité, 52
*F*isher, 6, 119
*F*luctuation, 55, 135
*F*onction
densité de probabilité, 39
de fréquence, 39
génératrice, 10

G

*G*auss, 22, 88
*G*rands nombres (loi des), 9
*G*rappes, 48

H

*H*asard, 46, 144, 148
*H*omogénéité, 195

*H*ypothèse
alternative, 140, 142
multiple, 139
nulle, 140
simple, 139

I
*I*ndépendance, 39
*I*négalité
de Bienaymé-Tchebychev, 9, 72
de Cramer-Rao, 75
*I*nférence, 39
*I*nformation, 39, 75
*I*ntervalle, 5
de confiance104, 110
de pari, 58
*I*nverse, 119

K - L
*K*hi-deux
distribution du, 17
table du, 18
test du, 192

*L*og-normale, 16
*L*oi
Binomiale, 12 , 77, 88
continue uniforme, 13
de Bernoulli, 12, 81, 88, 193
des grands nombres, 9
de Laplace-Gauss, 22, 88
exponentielle, 15, 88
de Fisher-Snedecor, 20, 162
du Khi-deux, 17
de Paréto, 22
de Poisson, 82, 88
de Student, 19, 164

uniforme, 13, 46

M

*M*ann-Whitney, 199
*M*arkov, 8
*M*aximum, 80, 86, 145
*M*eilleur, 102
*M*éthode, 83, 138, 160
aléatoires, 45
empiriques, 42
du maximum de vraisemblance, 78, 80, 194
*M*odèle, 12, 163
*M*oindres carrés, 78
*M*oments, 5, 11, 78, 151
*M*oyenne
d'un échantillon, 54, 56
test sur une, 149
de comparaisons de deux , 151

N

*N*eyman, 102, 136, 138
*N*iveau, 102
*N*ombre
au hasard, 47
loi faible(forte) des grands, 5

*N*on paramétrique, 136, 189
*N*ormale, 58, 88, 152
*N*otation, 54, 85
*N*ulle
hypothèse, 140, 160

O - P

*O*bservation, 160
*O*bservée, 196
fréquence, 197

valeur, 141, 166

*P*aramètre, 55, 85
de forme, 6
de dispersion, 4
*P*ari, 58
*P*as, 47
*P*aramétrique, 136, 201
*P*earson, 78, 138, 190
*P*oisson, 82
*P*olitz, 44
*P*ratique, 50, 102, 193
*P*récision, 38, 102
*P*robabilité
à priori, 138
à postériori, 138
inégale, 49
*P*roportion
estimation, 85
test sur, 142
test de comparaisons, 146
*P*uissance
d'un test, 137
test le plus puissant, 139

Q - R

*Q*uotas, 43, 48

*R*ao, 75
*R*ecensement, 39
*R*égion
critique, 136, 139, 149
d'acceptation, 136, 140
de rejet, 136
*R*ejet, 136, 164
*R*emise, 51

*R*ésiduelle, 161, 163
*R*isque, 141

S
*S*écurité, 102
*S*équences, 200
*S*euil, 104, 141
*S*ignificatif, 141
*S*imple, 139
*S*omme, 199
*S*ondage, 40
*S*ystématique, 47
*S*tatistique, 38, 53, 136, 166
*S*trate, 48
*S*tudent
    loi, 19, 164
    table, 20, 159
    test de, 164
    variable, 112

T
*T*ableau
    de contingence, 148
    de synthèse, 88, 166
*T*erminologie, 136
*T*héorème central limite, 9
*T*héorie, 135
*T*est
    baysien, 138
    bilatéral, 142
    d'ajustement, 190
    d'homogénéité, 195
    d'hypothèse, 138
    d'indépendance, 196
    de MCNEMAR, 148
    des rangs, 199
    de Fisher, 164

du χ^2, 190
des signes, 197
statistique, 136
UMP 143
Unilatéral, 145

U - V

*U*niforme, 13
*U*nilatéral, 145, 148
*U*tilisation, 17

*V*ariable, 7
aléatoire, 161, 191
centrée réduite, 4
*V*ariance,
d'un échantillon, 57
test sur une, 153
*V*raisemblance, 80, 84
fonction de, 80, 84
méthode du maximum de, 80, 194

TABLES Statistiques

(extraits)

I *T*able de la loi ***normale centrée réduite N(0, 1)***

II *T*able de la loi de χ^2

III *T*able de la loi de ***Student***

IV *T*able de la loi de ***Fisher-Snedecor***

V *T*able des ***nombres au hasard***

VI *T*ransformation de ***Fisher***

TABLE I
LOI NORMALE CENTREE REDUITE N(0,1)

Fonction de répartition $G(t) = P(T \leq t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$

t	0,00	0,01	0,02	0,03	0,040	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Pour $t < 0$, il faut prendre le complément de 1 de la valeur lue dans la table G(t):

$$G(-t) = 1 - G(t)$$

Table des grandes valeurs de t

t	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4	4,3
G (t)	0,99865	0,99904	0,99931	0,99952	0,99966	0,99977	0,99984	0,99993	0,99997	0,9999

TABLE II

TABLE DE DISTRIBUTION DE χ^2

Valeurs du χ^2 ayant la probabilité d'être dépassées

α	0,995	0,990	0,975	0,950	0,900	0,1	0,05	0,01	0,005	0,001
ν										
1	0,00004	0,00016	0,00098	0,00393	0,0158	2,7055	3,8414	6,6349	7,8794	10,828
2	0,01	0,0201	0,0506	0,1025	0,2107	4,6051	5,9914	9,2103	10,5966	13,816
3	0,0717	0,1148	0,2157	0,3518	0,5843	6,2513	7,8147	11,3449	12,8381	16,266
4	0,2069	0,2971	0,4844	0,7101	1,0636	7,7794	9,4877	13,2767	14,8602	18,467
5	0,4117	0,5543	0,8312	1,1454	1,6103	9,2363	11,0705	15,0863	16,7496	20,515
6	0,6757	0,8720	1,2373	1,6353	2,2041	10,6446	12,5916	16,8119	18,5476	22,458
7	0,9892	1,2390	1,6898	2,1673	2,8331	12,0170	14,0671	18,4753	20,2777	24,332
8	1,3444	1,6464	2,1797	2,7326	3,4895	13,3616	15,5073	20,0902	21,9550	26,125
9	1,7349	2,0879	2,7003	3,3251	4,1681	14,6837	16,9190	21,6660	23,5893	27,877
10	2,1558	2,5582	3,2469	3,9403	4,8651	15,9871	18,3070	23,2093	25,1882	29,588
11	2,6032	3,0534	3,8157	4,5748	5,5777	17,2750	19,6751	24,7250	26,7569	31,264
12	3,0738	3,5705	4,4037	5,2260	6,3038	18,5494	21,0261	26,2170	28,2995	32,909
13	3,5660	4,1069	5,0087	5,8918	7,0415	19,8119	22,3621	27,6883	29,8194	34,528
14	4,0746	4,6604	5,6287	6,5706	7,7895	21,0642	23,6848	29,1413	31,3193	36,123
15	4,6009	5,2293	6,2621	7,2609	8,5467	22,3072	24,7356	30,5779	32,8013	37,697
16	5,1422	5,8122	6,9076	7,9616	9,3122	23,5418	26,2962	31,9999	34,2672	39,252
17	5,6972	6,0477	7,5641	8,6717	10,0852	24,7690	27,5871	33,4087	35,7185	40,790
18	6,2648	7,0149	8,2307	9,3904	10,8649	25,9894	28,8693	34,8053	37,1564	42,312
19	6,8439	7,6327	8,9065	10,1170	11,6509	27,2036	30,1435	36,1908	38,5822	43,820
20	7,4338	8,2604	9,5908	10,8508	12,4426	28,4120	31,4104	37,5662	39,9968	45,315
21	8,0336	8,8972	10,2829	11,5913	13,2396	29,6151	32,6705	38,9321	41,4010	46,797
22	8,6427	9,5424	10,9823	12,3380	14,0415	30,8133	33,9244	40,2894	42,7956	48,268
23	9,2604	10,1956	11,6885	13,0905	14,8479	32,0069	35,1725	41,6384	44,1813	49,728
24	9,8862	10,8564	12,4011	13,8484	15,6587	33,1963	36,4151	42,9798	45,5585	51,179
25	10,5197	11,5240	13,1197	14,6114	16,4734	34,3816	37,6525	44,3141	46,9278	52,620
26	11,1603	12,1981	13,8439	15,3791	17,2919	35,5631	38,8852	45,6417	48,2899	54,052
27	11,8076	12,8786	14,5733	16,1513	18,1138	36,7412	40,1133	46,9630	49,6449	55,476
28	12,4613	13,5648	15,3079	16,9279	19,7677	37,9159	41,3372	48,2782	50,9933	56,892
29	13,1211	14,2565	16,0471	17,7083	20,5992	39,0875	42,5569	49,5879	52,3356	58,302
30	13,7867	14,9535	16,7908	18,4926	29,0505	40,2560	43,7729	50,8922	53,6720	59,703

Quand le ***nombre de degré de liberté*** est ***supérieur à 30***, on utilise la *table N (0,1)* avec $t = \sqrt{2\chi^2} - \sqrt{2\nu - 1}$

TABLE III
TABLE DE DISTRIBUTION DE t
(*loi de Student*)
Valeurs de t ayant la probabilité d'être dépassées en valeur absolue

P	0,9	0,8	0,70	0,50	0,30	0,20	0,1	0,05	0,01	0,001
ν										
1	0,158	0,325	0,510	1,000	1,963	3,078	6,314	12,706	63,657	636,619
2	0,142	0,289	0,445	0,816	1,386	1,886	2,920	4,303	9,925	31,598
3	0,137	0,277	0,424	0,765	1,250	1,638	2,353	3,182	5,841	12,929
4	0,134	0,271	0,414	0,741	1,190	1,533	2,132	2,776	4,604	8,610
5	0,132	0,267	0,408	0,727	1,156	1,476	2,015	2,571	4,032	6,869
6	0,131	0,265	0,404	0,718	1,134	1,440	1,943	2,447	3,707	5,959
7	0,130	0,263	0,402	0,711	1,119	1,415	1,895	2,365	3,499	5,408
8	0,130	0,262	0,399	0,706	1,108	1,397	1,860	2,306	3,355	5,041
9	0,129	0,261	0,398	0,703	1,100	1,383	1,833	2,262	3,250	4,781
10	0,129	0,260	0,397	0,700	1,093	1,372	1,812	2,228	3,169	4,587
11	0,129	0,260	0,396	0,697	1,088	1,363	1,796	2,201	3,106	4,437
12	0,128	0,259	0,395	0,695	1,083	1,356	1,782	2,179	3,055	4,318
13	0,128	0,259	0,394	0,694	1,079	1,350	1,771	2 ,160	3,012	4,221
14	0,128	0,258	0,393	0,692	1,076	1,345	1,761	2,145	2,977	4,140
15	0,128	0,258	0,393	0,691	1,074	1,341	1,753	2,131	2,947	4,073
16	0,128	0,258	0,392	0,690	1,071	1,337	1,746	2,120	2,921	4,015
17	0,128	0,257	0,392	0,689	1,069	1,333	1,740	2,110	2,898	3,965
18	0,127	0,257	0,392	0,688	1,067	1,330	1,734	2,101	2,878	3,922
19	0,127	0,257	0,391	0,688	1,066	1,328	1,729	2,093	2,861	3,883
20	0,127	0,257	0,391	0,687	1,064	1,325	1,725	2,086	2,845	3,850
21	0,127	0,257	0,391	0,686	1,063	1,323	1,721	2,080	2,831	3,819
22	0,127	0,256	0,390	0,686	1,061	1,321	1,717	2,074	2,819	3,792
23	0,127	0,256	0,390	0,685	1,060	1,319	1,714	2,069	2,807	3,767
24	0,127	0,256	0,390	0 ,685	1,059	1,318	1,711	2,064	2,797	3,745
25	0,127	0,256	0,390	0,684	1,058	1,316	1,708	2,060	2,787	3,725
26	0,127	0,256	0,390	0,684	1,058	1,315	1,706	2,056	2,779	3,707
27	0,127	0,256	0,389	0,684	1,057	1,314	1,703	2,052	2,771	3,690
28	0,127	0,256	0,389	0,683	1,056	1,313	1,701	2,048	2,763	3,674
29	0,127	0,256	0,389	0,683	1,055	1,311	1,699	2,045	2,756	3,649
30	0,127	0,256	0,389	0,683	1,055	1,310	1,697	2,042	2,750	3,656
40	0,126	0,255	0,388	0,681	1,050	1,303	1,684	2,021	2,704	3,551
120	0,127	0,254	0,386	0,677	1,041	1,289	1,658	1,980	2,617	3,373
∞	0,126	0,253	0,385	0,674	1,036	1,282	1,645	1,960	2,576	3,291

TABLE IV
TABLE DE DISTRIBUTION DE *F*
loi de Fisher-Snedecor
Valeurs de F ayant la probabilité d'être dépassées

	$\nu_1 = 1$		$\nu_1 = 2$		$\nu_1 = 3$		$\nu_1 = 4$		$\nu_1 = 5$	
	P 0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01
$\nu_1 = 2$										
1	161,4	4052	199,5	4999	215,7	5403	224,6	5625	230,2	5764
2	18,51	98,49	19,00	99,00	19,25	99,17	19,25	99,25	19,30	99,30
3	10,13	34,12	9,55	30,81	9,28	29,46	9,12	28,71	9,01	28,24
4	7,71	21,20	6,94	18,00	6,59	16,69	6,39	15,98	6,26	15,52
5	6,61	16,26	5,79	13,27	5,41	12,06	5,19	11,39	5,05	10,07
6	5,99	13,74	5,14	10,91	4,76	9,78	4,53	9,15	4,30	8,75
7	5,59	12,25	4,74	9,55	4,35	8,45	4,12	7,85	3,97	7,45
8	5,32	11,26	4,46	8,65	4,07	7,59	3,84	7,01	3,69	6,63
9	5,12	10,56	4,26	8,02	3,86	6,99	3,63	6,42	3,48	6,06
10	4,96	10,04	4,10	7,56	3,71	6,55	3,48	5,99	3,33	5,64
11	4,84	9,65	3,98	7,20	3,59	6,22	3,36	5,67	3,20	5,32
12	4,75	9,33	3,88	6,93	3,49	5,95	3,26	5,41	3,11	5,06
13	4,67	9,07	3,80	6,70	3,41	5,74	3,18	5,20	3,02	4,89
14	4,60	8,86	3,74	6,51	3,34	5,56	3,11	5,03	2,96	4,69
15	4,54	8,68	3,68	6,36	3,29	5,42	3,06	4,89	2,90	4,56
16	4,49	8,53	3,63	6,23	3,24	5,29	3,01	4,77	2,85	4,44
17	4,45	8,40	3,59	6,11	3,20	5,18	2,96	4,67	2,81	4,34
18	4,41	8,28	3,55	6,01	3,16	5,09	2,93	4,58	2,77	4,25
19	4,38	8,18	3,52	5,93	3,13	5,01	2,90	4,50	2,74	4,17
20	4,35	8,10	3,49	5,85	3,10	4,94	2,87	4,43	2,71	4,10
21	4,32	8,02	3,47	5,78	3,07	4,87	2,84	4,37	2,68	4,04
22	4,30	7,94	3,44	5,72	3,05	4,82	2,82	4,31	2,66	3,99
23	4,28	7,88	3,42	5,66	3,03	4,76	2,80	4,26	2,64	3,94
24	4,26	7,82	3,40	5,61	3,01	4,72	2,78	4,22	2,62	3,90
25	4,24	7,77	3,38	5,57	2,99	4,68	2,76	4,18	2,60	3,86
26	4,22	7,72	3,37	5,53	2,98	4,64	2,74	4,14	2,59	3,82
27	4,21	7,68	3,35	5,49	2,96	4,60	2,73	4,11	2,57	3,78
28	4,20	7,64	3,34	5,45	2,95	4,57	2,71	4,07	2,56	3,75
29	4,18	7,60	3,33	5,42	2,93	4,54	2,70	4,04	2,54	3,73
30	4,17	7,56	3,32	5,39	2,92	4,51	2,69	4,02	2,53	3,70
40	4,08	7,31	3,23	5,18	2,84	4,31	2,61	3,83	2,45	3,51
120	3,92	6,85	3,07	4,79	2,68	3,95	2,45	3,48	2,29	3,17
∞	3,84	6,64	2,99	4,60	2,60	3,78	2,37	3,32	2,21	3,02

TABLE IV *(suite)*

TABLE DE DISTRIBUTION DE F

loi de Fisher-Snedecor

Valeurs de F ayant la probabilité d'être dépassées

	$\nu_1 = 1$		$\nu_1 = 2$		$\nu_1 = 3$		$\nu_1 = 4$		$\nu_1 = 5$	
P	*0,05*	0,01	0,05	0,01	0,05	0,01	0,05	0,01	0,05	0,01
$\nu_1 = 2$										
1	234,0	5859	238,9	5981	243,9	6106	249,0	6234	254,3	6366
2	19,33	99,33	19,36	99,36	19,41	99,42	19,45	99,46	19,50	99,50
3	8,94	27,91	8,84	27,49	8,74	27,05	8,64	26,60	8,53	26,12
4	16,16	15,21	6,04	14,80	5,91	14,37	5,77	13,93	5,63	13,46
5	4,95	10,67	4,82	10,27	4,68	9,89	4,53	9,47	4,36	9,02
6	4,28	8,47	4,15	8,10	4,00	7,72	3,84	7,31	3,67	6,88
7	3,87	7,19	3,73	6,84	3,57	6,47	3,41	6,07	3,23	5,65
8	3,58	6,37	3,44	6,03	3,28	5,67	3,12	5,28	2,93	4,86
9	3,37	5,80	3,23	5,47	3,07	5,11	2,90	4,73	2,71	4,31
10	3,22	5,39	3,07	5,06	2,91	4,71	2,74	4,33	2,54	3,91
11	3,09	5,07	2,95	4,74	2,79	4,40	2,61	4,02	2,40	3,60
12	3,00	4,82	2,85	4,50	2,69	4,16	2,50	3,78	2,30	3,36
13	2,92	4,62	2,77	4,30	2,60	3,96	2,42	3,59	2,21	3,16
14	2,85	4,46	2,70	4,14	2,53	3,80	2,35	3,43	2,13	3,00
15	2,79	4,32	2,64	4,00	2,48	3,67	2,29	3,29	2,07	2,87
16	2,74	4,20	2,59	3,89	2,42	3,55	2,24	3,18	2,01	2,75
17	2,70	4,10	2,55	3,79	2,38	3,45	2,19	3,08	1,96	2,65
18	2,66	4,01	2,51	3,71	2,34	3,37	2,15	3,00	1,92	2,57
19	2,63	3,94	2,48	3,63	2,31	3,30	2,11	2,92	1,88	2,49
20	2,60	3,87	2,45	3,56	2,28	3,23	2,08	2,86	1,84	2,42
21	2,55	3,81	2,42	3,51	2,25	3,17	2,05	2,80	1,81	2,36
22	2,57	3,76	2,40	3,45	2,23	3,12	2,03	2,75	1,78	2,31
23	2,53	3,71	2,38	3,41	2,20	3,07	2,00	2,70	1,76	2,26
24	2,51	3,67	2,36	3,36	2,18	3,03	1,98	2,66	1,73	2,21
25	2,49	3,63	2,34	3,32	2,16	2,99	1,96	2,62	1,71	2,17
26	2,47	3,59	2,32	3,29	2,15	2,96	1,95	2,58	1,69	2,13
27	2,46	3,56	2,30	3,26	2,13	2,93	1,93	2,55	1,67	2,10
28	2,44	3,53	2,29	3,23	2,12	2,90	1,91	2,52	1,65	2,06
29	2,43	3,50	2,28	3,20	2,10	2,87	1,90	2,49	1,64	2,03
30	2,42	3,47	2,27	3,17	2,09	2,84	1,89	2,47	1,62	2,01
40	2,34	3,29	2,18	2,99	2,00	2,66	1,79	2,29	1,51	1,80
120	2,17	2,96	2,01	2,66	1,83	2,34	1,61	1,95	1,25	1,38
∞	2,09	2,80	1,94	2,51	1,75	2,18	1,52	1,79	1,00	1,00

TABLE V
TABLE DES NOMBRES AU HASARD

12651	61646	11769	75109	86996	97669	25757	32535
81769	74436	02630	72310	45049	18029	07469	42341
36737	98863	77240	76251	00654	64688	09343	70278
82861	54371	76610	94934	72748	44124	05610	53750
21325	15732	24127	37341	09723	63529	73977	95218
74146	47887	62463	23045	41490	07954	22597	60012
90759	64410	54179	66075	61051	75385	51378	08360
55683	98078	02238	91540	21219	17720	87817	41705
79686	17969	76061	83748	55920	83612	41540	86492
70333	00201	86201	69716	78185	62154	77930	67663
14042	53536	07779	04157	41172	36473	42123	43929
59911	08256	06596	48416	69770	68797	56080	14223
62368	62623	62742	14891	39247	52242	98832	69533
57529	97751	54976	48957	74599	08759	78494	52785
15469	90574	78033	66885	13936	42117	71831	22961
18625	23674	53850	32827	81647	80820	00420	63555
74626	68394	88562	70745	23701	45630	65891	58220
11119	16519	27384	90199	79210	76965	99546	30323
41101	17336	48951	53674	17880	45260	08575	49321
32123	91576	84221	78902	82010	30847	62329	63898
26091	68409	69704	82267	14751	13151	93115	01437
67680	79790	48462	59278	44185	29616	76531	19589
15184	19260	14073	07026	25264	08388	27182	22557
58010	45039	57181	10238	36874	28546	37444	80824
56425	53996	86245	32623	78858	08143	60377	42925
82630	84066	13592	60642	17904	99718	63432	88642
14927	40909	23900	48761	44860	92467	31742	87142
23740	22505	07489	85986	74420	21744	97711	36648
32990	97446	03711	63824	07953	85965	87089	11687
05310	24058	91946	78437	34365	82469	12430	84754
21839	39937	27534	88913	49055	19218	47712	67677
08833	42549	93981	94051	28382	83725	72643	64233
58336	11139	47479	00931	91560	95372	97642	33856
62032	91144	75478	47431	52726	30289	42411	91886

TABLE VI
TRANSFORMATION DE FISHER

$$z' = \frac{1}{2} \ln \frac{1 + r}{1 - r}$$

0,00	0,000	0,25	0,255	0,50	0,549	0,75	0,973
0,01	0,010	0,26	0,266	0,51	0,563	0,76	0,996
0,02	0,020	0,27	0,277	0,52	0,576	0,77	1,020
0,03	0,030	0,28	0,288	0,53	0,590	0,78	1,045
0,04	0,040	0,29	0,299	0,54	0,604	0,79	1,071
0,05	0,050	0,30	0,310	0,55	0,618	0,80	1,099
0,06	0,060	0,31	0,321	0,56	0,633	0,81	1,127
0,07	0,070	0,32	0,332	0,57	0,648	0,82	1,157
0,08	0,080	0,33	0,343	0,58	0,662	0,83	1,188
0,09	0,090	0,34	0,354	0,59	0,678	0,84	1,221
0,10	0,100	0,35	0,365	0,60	0,693	0,85	1,256
0,11	0,110	0,36	0,377	0,61	0,709	0,86	1,293
0,12	0,121	0,37	0,388	0,62	0,725	0,87	1,333
0,13	0,131	0,38	0,400	0,63	0,741	0,88	1,376
0,14	0,141	0,39	0,412	0,64	0,758	0,89	1,422
0,15	0,151	0,40	0,424	0,65	0,775	0,90	1,472
0,16	0,161	0,41	0,436	0,66	0,793	0,91	1,528
0,17	0,172	0,42	0,448	0,67	0,811	0,92	1,589
0,18	0,182	0,43	0,460	0,68	0,829	0,93	1,658
0,19	0,192	0,44	0,472	0,69	0,848	0,94	1,738
0,20	0,203	0,45	0,485	0,70	0,867	0,95	1,832
0,21	0,213	0,46	0,497	0,71	0,887	0,96	1,946
0,22	0,224	0,47	0,510	0,72	0,908	0,97	2,092
0,23	0,234	0,48	0,523	0,73	0,929	0,98	2,298
0,24	0,245	0,49	0,536	0,74	0,950	0,99	2,647

Impression : SAGEP - La Kasbah - Tunis

ISBN 9973-41-058-0
Tous droits réservés à l'auteur et à ses ayants droits *Septembre / 2001*

Rejeb HAJI enseignant à l'Institut Supérieur de Gestion de Tunis, a également dispensé des cours à l'Ecole Nationale d'Administration et à la faculté de droit et des sciences économiques de Tunis. Il est titulaire d'un doctorat d'Etat ès-sciences économiques de la faculté d'économie appliquée d'Aix Marseille, d'un doctorat de 3ème cycle en statistiques de l'université de Paris VI, d'un diplôme de l'Institut d'Administration des Entreprises de Paris (IAE) et du diplôme de l'Institut de Défense Nationale (4ème promotion).

Il a suivi des sessions de formation à Harvard University, à l'Université de Belgrade et à l'Institut International d'Administration Publique de Paris.

En détachement auprès de l'Administration, il a dirigé la Direction des Etudes et de la Planification du Ministère de l'Equipement et a assumé différentes fonctions dans plusieurs cabinets ministériels (Ministère de la Jeunesse et des Sports, Ministère délégué auprès du Premier Ministre, Ministère de la santé publique).

Il a publié de nombreux articles et communications scientifiques notamment dans la Revue Tunisienne des Sciences Sociales et dans la Revue Tunisienne d'Economie et de Gestion.

Il est titulaire de plusieurs décorations dont celle d'officier de l'ordre de l'indépendance, de l'ordre de la république et de la médaille de la jeunesse. Marié, il est père de deux enfants.

Le présent manuel est le fruit de ces expériences diversifiées et d'une réflexion approfondie pour l'enseignement des Statistiques au service de la gestion.

Cet ouvrage complète le manuel **«Statistique & Probabilités»**. Comme le précédent, il privilégie l'application et limite le poids de la théorie. On y trouve, en **six chapitres** avec des cas pratiques corrigés une introduction à **l'inférence statistique : estimation de paramètres- Sondage- tests d'hypothèses- tests du Khi-deux «χ^2»**

Ce livre s'adresse aux étudiants du **second cycle universitaire** et à tous ceux qui désirent confronter «les faits et les modèles».

ISBN 9973-41-058-0

Prix : 13 DT